

AD-A034 465

WISCONSIN UNIV MADISON MATHEMATICS RESEARCH CENTER

F/G 12/1

DIFFERENCE METHODS FOR STIFF ORDINARY DIFFERENTIAL EQUATIONS.(U)

NOV 76 H KREISS

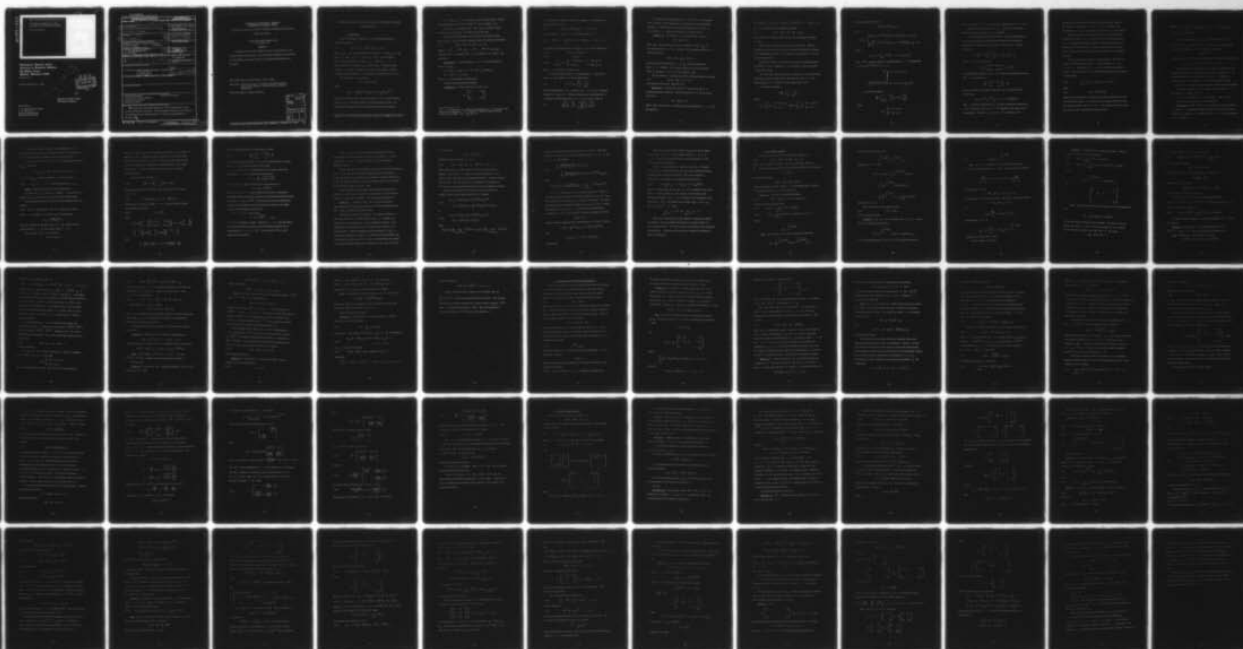
DAAG29-75-C-0024

UNCLASSIFIED

MRC-TSR-1699

NL

1 of 1
ADA034465



END

DATE
FILMED
2 - 77

ADA034465

MRC Technical Summary Report # 1699

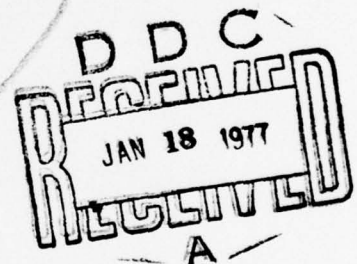
DIFFERENCE METHODS FOR STIFF
ORDINARY DIFFERENTIAL EQUATIONS

Heinz-Otto Kreiss

Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, Wisconsin 53706

November 1976

(Received September 3, 1976)



Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 1699 ✓	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DIFFERENCE METHODS FOR STIFF ORDINARY DIFFERENTIAL EQUATIONS. ✓		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
7. AUTHOR(s) Heinz-Otto Kreiss		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of 610 Walnut Street Wisconsin Madison, Wisconsin 53706 ✓		8. CONTRACT OR GRANT NUMBER(s) DAAG29-75-C-0024 ✓
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Technical summary rept.		12. REPORT DATE November 1976
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. 1269p. 14 MRC-TSR-1699		13. NUMBER OF PAGES 66
17. DISTRIBUTION STATEMENT (of the Abstract entered in Block 20, if different from Report)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
18. SUPPLEMENTARY NOTES		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Numerical analysis Singular perturbation Ordinary differential equations Initial value problems Asymptotic theory		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Consider the initial value problem for a first order system of stiff ordinary differential equations. The smoothness properties of its solutions are investigated and a general theory for difference approximations is developed. *		

DD FORM 1473

EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

DIFFERENCE METHODS FOR STIFF ORDINARY DIFFERENTIAL EQUATIONS

Heinz-Otto Kreiss

Technical Summary Report # 1699
November 1976


ABSTRACT

Consider the initial value problem for a first order system of stiff ordinary differential equations. The smoothness properties of its solutions are investigated and a general theory for difference approximations is developed.

AMS (MOS) Subject Classifications: 65L05, 34E15

Key Words: Numerical analysis, Ordinary differential equations,
Initial value problems, Asymptotic theory, Singular
perturbation

Work Unit Number 1 (Applied Analysis)

ACCESSION FOR	
NTIS	White Section <input checked="" type="checkbox"/>
DDC	Buff Section <input type="checkbox"/>
UNANNOUNCED	<input type="checkbox"/>
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	AVAIL. and/or SPECIAL
	

DIFFERENCE METHODS FOR STIFF ORDINARY DIFFERENTIAL EQUATIONS

Heinz-Otto Kreiss

1. Introduction.

One of the simplest examples of a stiff differential equation is the scalar equation

$$(1.1) \quad dy/dt = a_{11}y + be^{dt}, \quad y(0) = y_0, \quad t \geq 0.$$

Here a_{11}, b, d are complex valued constants with $\operatorname{Re} a_{11} \leq 0$ and $\operatorname{Re} d \leq 0$. Also $|a_{11}| \gg 1$ but b/a_{11} and d are of moderate size. A typical example is given by $a_{11} = -10^4$, $b/a_{11} = 1$, $d = i$. The expression "of moderate size" is rather vague. It depends on the stepsize k one wants to employ in a numerical calculation. Often it is satisfactory to say that K is a constant of moderate size if $Kk \leq 0.1$.

The solution of (1.1) is given by

$$y(t) = y_1(t) + y_2(t),$$

where

$$y_1(t) = -\frac{b}{a_{11} - d} e^{dt}, \quad y_2(t) = \left(y_0 + \frac{b}{a_{11} - d}\right) e^{a_{11}t}.$$

Thus the solution consists of the forced solution $y_1(t)$ and the transient solution $y_2(t)$. The forced solution is smooth and varies slowly. For the transient solution we have two fundamentally different situations.

1) $\text{Re } a_{11} \ll -1$. In this case $y_2(t)$ decays rapidly. Outside a boundary layer of order $O(|a|^{-1} |\log |a||)$ we can neglect $y_2(t)$.

2) $\text{Re } a_{11}$ is of moderate size. Now $y_2(t)$ oscillates rapidly but does not decay quickly. We shall not treat that case.

This division of the solution into a slowly and a rapidly varying part is typical. Let us consider a general system with constant coefficients

$$(1.2) \quad dy/dt = Ay + F(t), \quad y(0) = y_0, \quad t \geq 0,$$

where $y = (y^{(1)}, \dots, y^{(n)})'$ and $F = (F^{(1)}, \dots, F^{(n)})'$ are vector functions with n components and A is a constant $n \times n$ matrix. We shall make

Assumption 1.1. The eigenvalues λ_j of A can be divided into two sets M_1, M_2 .

1) $\lambda_j \in M_1$ if $\text{Re } \lambda_j \ll -1$ and $|\text{Im } \lambda_j| \leq \rho |\text{Re } \lambda_j|$.

2) $\lambda_j \in M_2$ if $|\lambda_j| \leq \frac{1}{2} C$.

Here ρ, C are constants of moderate size.

We need also the following concepts:

Definition 1.1. A matrix function $A = A(t), t \geq 0$

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \cdot & \cdot & \cdot \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

1) If y is a vector then y' denotes its transpose and y^* its adjoint. The vector norm is defined by $|y| = \max |y^{(i)}|$. Similar notations hold for matrices, for example $|A| = \sup_y |Ay|/|y|$.

is called negative dominant if there is a constant ρ of moderate size such that for all $t \geq 0$

$$(1.3) \quad |\operatorname{Im} a_{ii}| \leq \rho |\operatorname{Real} a_{ii}|, \quad i = 1, 2, \dots, n;$$

and a constant δ with $0 < \delta \leq 1$ such that for all $t \geq 0$

$$(1.4) \quad \operatorname{Real} a_{ii} < -1, \quad \sum_{\substack{j=1 \\ i \neq j}}^n |a_{ij}| < -(1 - \delta) \operatorname{Real} a_{ii}, \quad i = 1, 2, \dots, n.$$

It is called essentially negative dominant if (1.3), (1.4) are replaced by

$$(1.3a) \quad |\operatorname{Im} a_{ii}| \leq \rho |\operatorname{Real} a_{ii}| + c$$

$$(1.4a) \quad \operatorname{Real} a_{ii} < c$$

$$(1.4b) \quad \sum_{j=1, j \neq i}^n |a_{ij}| \leq \begin{cases} -(1 - \delta) \operatorname{Real} a_{ii} + c & \text{if } \operatorname{Real} a_{ii} \leq 0 \\ c - \operatorname{Real} a_{ii} & \text{if } \operatorname{Real} a_{ii} > 0 \end{cases}$$

where c is a constant of moderate size.

It is well known that there is a transformation S such that the matrix A of the system (1.2) can be transformed to

$$(1.5) \quad \tilde{A} = S^{-1}AS = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix},$$

where the eigenvalues of A_{ii} belong to M_i , $i = 1, 2$. A_{11} is negative dominant and $|A_{22}| \leq 2c$. Without restriction we can assume that A is already in blockdiagonal form (1.5), i.e. (1.2) can be written as

$$(1.6) \quad \frac{d}{dt} \begin{pmatrix} y^I \\ y^{II} \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} y^I \\ y^{II} \end{pmatrix} + \begin{pmatrix} F^I \\ F^{II} \end{pmatrix}.$$

Of course, we cannot guarantee that $|S| + |S^{-1}|$ is of moderate size. However, if this expression is large then the original dependent variables y had not been correctly chosen and we can consider S as a scaling of y . In the next section we are going to prove

Theorem 1.1. The solution of (1.6) can be written in the form

$$y(t) = y_S(t) + v(t) .$$

Here $y_S(t)$ and its derivatives can be estimated by $A_{11}^{-1} F^I$, A_{22} , F^{II} and their derivatives. $v(t)$ is a solution of the homogenous system (1.6) with initial values

$$v^I(0) = y^I(0) - y_S^I(0), \quad v^{II}(0) = 0 .$$

By assumption the eigenvalues λ_j of A_{11} have the property that $\text{Real } \lambda_j \ll -1$. Therefore $v(t)$ decays rapidly and, outside a boundary layer, the solution y of (1.6) is as smooth as $y_S(t)$.

The last theorem is a special case of the following theorem for systems

$$(1.7) \quad dy/dt = A(t)y + F(t), \quad t \geq 0 ,$$

with variable coefficients. (See also [2].)

Theorem 1.2. Consider the system (1.7) and assume that A is essentially negative dominant. Then the solutions of (1.7) can be written in the form

$$y(t) = y_S(t) + v(t) ,$$

where $y_S(t)$ and its first p derivatives can be estimated by c , δ , ρ and the functions

$$(1.8) \quad \min(|a_{ii}|^{-1}, 1) d^v a_{ij} / dt^v, \min(|a_{ii}|^{-1}, 1) d^v F^{(i)} / dt^v, \quad v = 0, 1, 2, \dots, p.$$

$v(t)$ is the solution of the homogenous equation

$$(1.9) \quad dv/dt = A(t)v, \quad v(0) = y(0) - y_S(0),$$

which away from a boundary layer has the same smoothness properties as $y_S(t)$.

Assume that the functions (1.8) are of moderate size. Then the boundary layer can always be resolved by making a logarithmic stretching of the independent variable t . If we do this, the solutions of (1.7) will be smooth everywhere.

It should be pointed out, that we do not make any assumption that the number of "large" eigenvalues of A is constant. Thus we are able to treat turningpoints.

One might think that the conditions of Theorem 1.2 are too restrictive. We shall give three examples which show that this is not so. In all these examples $\epsilon > 0$ denotes a small constant and $t \geq -1$.

1) Consider the system

$$\epsilon \frac{dy}{dt} = \begin{pmatrix} 0 & 0 \\ a_{21} & a_{22} \end{pmatrix} y,$$

where

$$a_{21} = \begin{cases} 1 & \text{for } -1 \leq t \leq 0 \\ e^{-t^2/\epsilon} & \text{for } t > 0, \end{cases} ; a_{22}(t) = \begin{cases} -1 & \text{for } -1 \leq t \leq 0 \\ -e^{-t/\epsilon} & \text{for } t > 0. \end{cases}$$

An easy calculation gives us

$$y^{(1)}(t) \equiv y^{(1)}(-1), \quad t \geq -1,$$

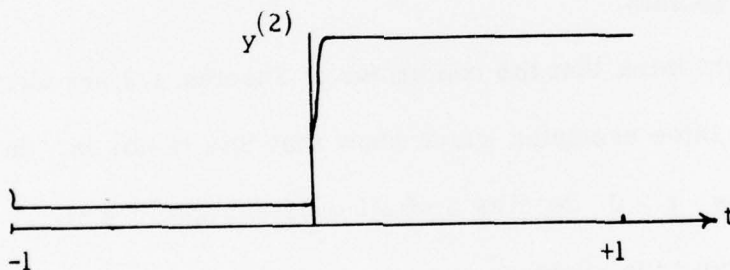
and

$$y^{(2)}(t) = \begin{cases} y^{(1)}(-1) + e^{-(t+1)/\epsilon} (y^{(2)}(-1) - y^{(1)}(-1)) & \text{for } -1 \leq t \leq 0, \\ \frac{1}{\sqrt{\epsilon}} \int_0^{t/\sqrt{\epsilon}} e^{-\sigma(t, \eta) - \eta^2} d\eta y^{(1)}(-1) + e^{-\sigma(t, 0)} y^{(2)}(0) & \text{for } t > 0, \end{cases}$$

where

$$\sigma(t, \eta) = \epsilon^{-1} \int_{\eta\sqrt{\epsilon}}^t e^{-\xi/\epsilon} d\xi, \quad \text{i.e. } 0 \leq \sigma(t, \eta) \leq 1.$$

Thus $y^{(2)}(t)$ changes rapidly in a neighbourhood of $t = 0$ and becomes of order $\mathcal{O}(\epsilon^{-1/2} |y^{(1)}(-1)|)$ for $t > 0$.



2) Consider the system

$$\frac{dy}{dt} = \begin{pmatrix} 0 & \epsilon^{-2} \\ a_{21}(t) & \epsilon^{-1} \end{pmatrix} y, \quad y(-1) = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

where

$$a_{21} = \begin{cases} 0 & \text{for } t \leq 0, \\ -t^2 & \text{for } t \geq 0. \end{cases}$$

Now $y(t) \equiv (1, 0)'$ for $t \leq 0$ and converges rapidly to zero for $t > O(\sqrt{\epsilon})$.

Observe that the coefficients a_{ij} or $a_{ij}^{-1} da_{ij}/dt$ are bounded.

Thus the diagonal dominance of A is important.

3) One might think that one has only to control the eigenvalues of A and their variation to be sure that the solution of the differential equations is of moderate size and smooth. This is not so. Consider the system

$$\epsilon \frac{dy}{dt} = U^*(t) \begin{pmatrix} -1 & \eta \\ 0 & -1 \end{pmatrix} U(t) y = A(t)y, \quad y(0) = y_0,$$

where $\eta > 0$ is a constant and

$$U(t) = \begin{pmatrix} \cos \alpha t & \sin \alpha t \\ -\sin \alpha t & \cos \alpha t \end{pmatrix}, \quad \alpha = \text{const.},$$

is a unitary transformation. Let $v = Uy$ denote new dependent variables.

Then v is the solution of

$$\frac{dv}{dt} = \begin{pmatrix} -1/\epsilon & \eta/\epsilon - \alpha \\ \alpha & -1/\epsilon \end{pmatrix} v = \tilde{B}v.$$

This is a system with constant coefficients and its general solution

has the form

$$v(x) = e^{\kappa_1 x} \ell_1 + e^{\kappa_2 x} \ell_2, \quad \kappa_j = -\epsilon^{-1} \pm \sqrt{\alpha(\eta/\epsilon - \alpha)}.$$

Here κ_j are the eigenvalues and ℓ_j are the corresponding eigenvectors of \tilde{B} . As long as $\alpha(\eta/\epsilon - \alpha) < \epsilon^{-2}$ the solutions of the system decay exponentially. However, if $\alpha(\eta/\epsilon - \alpha) > \epsilon^{-2}$ then there is one

exponentially increasing and one exponentially decreasing solution.

This happens, for example, if $\alpha = 2$, $\eta = \varepsilon^{-1} > 2$. Observe, that the eigenvalues of $A(t)$ are constant and that $U(t)$ is a slow rotation.

The above examples show that one needs to be cautious when solving stiff systems. It is, of course, not necessary that the system (1.7) be essentially diagonally dominant. It is sufficient that one can transform it to this normal form. In section three we shall show that this is always possible using an adequate stretching of the independent variables.

We shall now discuss difference approximations. We divide the t -axis into subintervals of length k , define gridpoints $t_v = vk$, $v = 0, 1, 2, \dots$ and denote by $u_v = u(t_v)$ vector functions defined on the grid. We approximate (1.7) by a multistep-method which is generated from an identity

$$(1.10) \quad \sum_{j=-1}^r \alpha_j y(t_{v-j}) + k\beta_j dy(t_{v-j})/dt = k\psi(t_v),$$

where

$$(1.11) \quad \psi(t_v) = \mathcal{O}(k^{r-1} d^r y/dt^r)$$

denotes the truncation error. Observe that (1.10) has nothing to do with the differential equation. It is valid for all sufficiently smooth functions. Only when we replace dy/dt by $Ay + F$ the differential equation enters and we obtain the corresponding multistep-method,

$$(1.12) \quad L[u] \equiv \sum_{j=-1}^r (\alpha_j I + k\beta_j A(t_{v-j})) u_{v-j} = kG_v, \quad G_v = - \sum_{j=-1}^r \beta_j F(t_{v-j}).$$

(1.12) can also be written as

$$(\alpha_{-1} I + k\beta_{-1} A(t_{v+1})) u_{v+1} = - \sum_{j=0}^r (\alpha_j I + k\beta_j A(t_{v-j})) u_{v-j} + kG_v.$$

We assume that $\alpha_{-1} = 1$ and that $\beta_{-1} \leq 0$. Then $(\alpha_{-1} I + k\beta_{-1} A(t_{v+1}))^{-1}$ exists if the eigenvalues λ of A satisfy the inequality

$$k\beta_{-1} \operatorname{Re} \lambda > -1.$$

Now the solution of (1.12) is uniquely determined if we specify initial values

$$u_{\mu} = y_{\mu} + \varphi_{\mu}, \quad |\varphi_{\mu}| = O(k^{r-1}), \quad \mu = 0, 1, 2, \dots, r.$$

The approximation is only useful if it is stable. However, here we cannot define stability by describing the behavior of a method as $k \rightarrow 0$, because we want to use it in the case that $k|A| \gg 1$. Instead we consider classes \mathcal{K} of problems (1.7), and define uniform stability.

Let Ω be a domain in the complex plane with the following properties

1) There is a constant C_1 (of moderate size) such that $z \in \Omega$ and $\operatorname{Re} z \geq 0$ implies $|z| \leq C_1$.

2) If $z \in \Omega$ then also $\sigma z \in \Omega$ for all real σ with $0 \leq \sigma \leq 1$.

We shall consider the class $\mathcal{M}(\Omega)$ defined by

Definition 1.2. $\mathcal{M}(\Omega) = \mathcal{M}(\Omega, c, \rho, \delta, K)$ denotes the class of problems (1.7) where A is essentially negative dominant, the functions (1.8) are bounded by a constant K (of moderate size) for $p = 1$ and the eigenvalues λ of A belong to Ω .

Stability is defined almost as usual. We assume that one can solve (1.12) for u_{v+1} boundedly and that the solutions of (1.12) increase at most slowly. The only difference is that all the constants involved are independent of the particular problem but hold for the whole class.

Definition 1.3. Consider a class \mathcal{K} of problems (1.7). A multi-step-method (1.12) is stable for the class if there are constants K_J, K_S, γ_S and $k^{(0)} > 0$ such that for all problems in \mathcal{K} , all t , and all k with $0 \leq k \leq k_0$

$$(1.13) \quad |(\alpha_{-1}I + k\beta_{-1}A(t))^{-1}| \leq K_J,$$

and the solution of the homogenous equation

$$(1.14) \quad L[v] \equiv \sum_{j=-1}^r (\alpha_j I + k\beta_j A(t_{v-j}))v_{v-j} = 0$$

satisfy the estimate

$$(1.15) \quad |v(t_\mu)| \leq K_S e^{\gamma_S(t_\mu - t_\nu)} \sum_{j=0}^r |v(t_{v-j})|, \quad t_\mu > t_\nu.$$

Already G. Dahlquist [1] has defined the stability for classes of problems. He considered the class η_0 of all scalar equations

$$(1.16) \quad dy/dt = \lambda y, \quad \lambda = \text{const. Real } \lambda \leq 0,$$

and proved that the trapezoidal rule is the "best method" (method of the highest order with the smallest coefficient of the truncation error) which

is stable for the class η_0 . Later O. Widlund [8] introduced the class

$\eta_1 = \eta_1(\rho)$ of all scalar equations (1.16) where λ satisfies the condition

$$(1.17) \quad \text{Real } \lambda \leq 0, \quad |\text{Im } \lambda| \leq \rho |\text{Real } \lambda|,$$

and showed that there are methods of higher order than two which are stable for such a class. There are now a large number of stable methods available. See for example [5], [9]. More generally, one can consider classes $\mathcal{N}(\Omega)$ of scalar equations (1.16) where λ belongs to a domain Ω of the above type.

We consider now the approximation (1.12) for the class $\mathcal{N}(\Omega)$. In this case the solutions can be given explicitly. They are of the form

$$(1.18) \quad u_\nu = \sum P_\mu(\nu) \kappa_\mu^\nu$$

where the κ_μ are the solutions of the characteristic equation

$$(1.19) \quad \sum_{j=-1}^r (\alpha_j + \lambda k \beta_j) \kappa^{r-j} = 0$$

and the P_μ are polynomials in ν of order one less than the multiplicity of κ_μ . We make the following

Assumption 1.2. 1) The approximation (1.12) is stable in a neighbourhood of $\lambda k = 0$ which means:

a) For $\lambda k = 0$ the solutions of (1.19) satisfy the condition

$$(1.20) \quad |\kappa_\mu(0)| \leq 1. \quad \text{If } |\kappa_\mu(0)| = 1 \text{ then its multiplicity is one.}$$

b) No weak instability occurs, i.e. in a neighbourhood of $\lambda k = 0$

the solutions $\kappa_\mu(\lambda k)$ with $|\kappa_\mu(0)| = 1$ are of the form

$$(1.21) \quad \kappa_\mu(\lambda k) = \kappa_\mu(0)(1 + g_\mu \lambda k) + o(\lambda k)^2 \quad \text{with } g_\mu > 0.$$

2) The approximation is stable in a neighbourhood of $\lambda k = \infty$.

(Dividing (1.19) by λk we can consider its solutions as functions of $(\lambda k)^{-1}$ and the stability is defined in the same way as above.)

3) Let d_1, d_2 with $0 < d_1 < d_2 < \infty$ be constants and denote by Ω_{d_1, d_2} the domain

$$z \in \Omega_{d_1, d_2} \text{ if } \operatorname{Re} z \leq 0 \text{ and } d_1 \leq |z| \leq d_2.$$

If $\lambda k \in \Omega \cap \Omega_{d_1, d_2}$ then the solutions κ_μ of (1.19) satisfy

$$(1.22) \quad |\kappa_\mu| < 1 - \tau, \quad \tau > 0 \text{ a constant depending on } d_1, d_2.$$

Remark. The last condition is often strengthened to:

For every fixed $d_1 > 0$ there is a constant τ such that (1.22) holds for all d_2 . In this case we call the method strongly stable.

As an example we consider methods which are generated from the identity

$$(1.23) \quad y(t_{\nu+1}) - \sigma k dy(t_{\nu+1})/dt = y(t_\nu) + (1 - \sigma) dy(t_\nu)/dt + k\psi(t_\nu),$$

where σ is a constant. (1.23) leads to the approximations

$$u_{\nu+1} = \frac{1 + k(1 - \sigma)\lambda}{1 - k\sigma\lambda} u_\nu,$$

which are unstable for any class $\eta_1(\rho)$ if $\sigma < \frac{1}{2}$; stable, but not strongly stable, if $\sigma = \frac{1}{2}$ and strongly stable if $\sigma > \frac{1}{2}$.

Now consider a class \mathcal{K} of problems

$$dy/dt = A(t)y + F$$

and assume that the class is such that the eigenvalues $\lambda(t)$ of every A belong to a set Ω . Approximate these problems by a multistep-method which is stable for the class $\mathcal{N}(\Omega)$. The main question which we want to discuss in Section 4 is whether the multistep-method is also stable for the class \mathcal{N} . In this generality the answer is no, as we shall demonstrate now.

Consider our third example, i.e.

$$(1.24) \quad \varepsilon \frac{dy}{dt} = U^*(t) \begin{pmatrix} -1 & \eta \\ 0 & -1 \end{pmatrix} U(t) y = A(t) y,$$

and approximate it by the multistep-method generated by (1.23) with $\sigma \geq \frac{1}{2}$, i.e.

$$(1.25) \quad (I - k\sigma A(t_{\nu+1}))u_{\nu+1} = (I + k(1 - \sigma)A(t_{\nu}))u_{\nu}.$$

Introduce into (1.25) new variables $w = Uu$. Then we obtain a system with constant coefficients

$$(1.26) \quad w_{\nu+1} = Bw_{\nu}$$

where

$$B = \left(I - \frac{\sigma k}{\varepsilon} \begin{pmatrix} -1 & \eta \\ 0 & -1 \end{pmatrix} \right)^{-1} \begin{pmatrix} \cos \alpha k_{\nu} & \sin \alpha k_{\nu} \\ -\sin \alpha k_{\nu} & \cos \alpha k_{\nu} \end{pmatrix} \left(I + (1 - \sigma)k_{\nu} \begin{pmatrix} -1 & \eta \\ 0 & -1 \end{pmatrix} \right) =$$

$$\begin{pmatrix} 1 & \tilde{\eta} \sigma \\ 0 & 1 \end{pmatrix} \left(I + \alpha k \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \mathcal{O}(\alpha k)^2 \right) \begin{pmatrix} \mu & \tilde{\eta}(1 - \sigma) \\ 0 & \mu \end{pmatrix},$$

with

$$\tilde{\eta} = \frac{k\eta}{\varepsilon} \left/ \left(1 + \frac{\sigma k}{\varepsilon} \right) \right., \quad \mu = \left(1 - \frac{(1 - \sigma)k}{\varepsilon} \right) \left/ \left(1 + \frac{\sigma k}{\varepsilon} \right) \right. .$$

(1.26) is an approximation of the homogenous system

$$(1.27) \quad \varepsilon \frac{dv}{dt} = \begin{pmatrix} -1 & \eta - \alpha\varepsilon \\ \alpha\varepsilon & -1 \end{pmatrix} v = \tilde{B}v$$

and we know that the solutions of (1.27) decay exponentially as long as $\varepsilon\alpha\eta = \tau < 1$. We shall now consider three cases.

1) $\sigma = \frac{1}{2}$, $\varepsilon \ll k$. A somewhat tedious calculation shows that eigenvalues κ of B are approximatively given by

$$\kappa \approx -1 + \frac{4\varepsilon}{k} \pm \sqrt{8\alpha\eta\varepsilon - \alpha^2 k^2}.$$

If $\alpha > 0$ and $\varepsilon = \frac{1}{4}\alpha k^2$ then we get for the negative root

$$\kappa \approx -1 + \alpha k(1 - \sqrt{2\eta - 1}),$$

and $|\kappa| > 1$ for $\eta > 1$. Thus the approximation has an exponential increasing solution, though the solution of the differential equation decays as long as $\eta < 1/\varepsilon\alpha = 4/\alpha^2 k^2$. Furthermore the increasing exponential solution can grow arbitrary fast and therefore the method is not stable for the class of problems of type (1.24):

2) $\sigma = 1$, $\varepsilon \ll k$. Now we get

$$\kappa \approx -\frac{\tau}{2} \pm \sqrt{\frac{1}{4}\tau^2}, \quad \tau = \varepsilon\alpha\eta.$$

If $|\tau| < 1$ the method is stable. However, if $\tau < -1$, then the method is not stable though the solutions of the differential equations decay exponentially as $t \rightarrow \infty$. ($\eta = O(\varepsilon^{-1})$ and the matrix \tilde{B} is far from being negative dominant.)

3) $\frac{1}{2} < \sigma \leq 1$, $\epsilon \ll k$. In this case the approximation has an exponentially increasing solution for $|\epsilon \alpha \eta| \geq \tau_0$ where τ_0 is some constant with $0 < \tau_0 < 1$. Again the matrix \tilde{B} is far from being negative dominant.

In some sense the most disturbing result is the effect on the trapezoidal rule ($\sigma = \frac{1}{2}$) because the difference approximation has exponential increasing solutions for examples which cannot be considered pathological at all. This kind of behavior is typical for methods which are only stable and not strongly stable for a class $\mathcal{M}(\Omega)$.

The last example shows that one cannot decide the stability of a method for a given class of problems by just looking at scalar equations. One has to impose other conditions. The main result of Section 4 is

Theorem 1.3. Consider a class $\mathcal{M}(\Omega)$ and assume that the approximation (1.12) satisfies assumption 1.2 for the corresponding class $\mathcal{M}(\Omega)$. Then it is stable for the class $\mathcal{M}(\Omega)$.

We shall now derive error estimates. Consider the problem (1.7) and assume that it has a smooth solution $y(t)$, i.e. a number of its derivatives are of moderate size. If the conditions of theorem 1.2 are fulfilled and the functions (1.8) are of moderate size, then this is no real restriction. Either we change the initial conditions to $y_S(0)$ or we perform a logarithmic scaling of t which resolves the boundary layer so that also the solution of (1.9) is smooth. We approximate the problem (1.7) by the multistep-method (1.12). Introducing $y(t)$ into (1.12) gives

us, using (1.10)

$$L[y] = kG_\nu + k\psi(t_\nu) .$$

Therefore we get for the error $w = y - u$

$$(1.26) \quad L[w] = k\psi(t_\nu), \quad w_\mu = \varphi_\mu = O(k^{r-1}), \quad \mu = 0, 1, \dots, r .$$

If the approximation is stable, then (1.26) gives us the usual error estimate. Observe that ψ depends only on the derivatives of $y(t)$ and not on the coefficients of the differential equation. Thus the stiffness does not enter. Thus we need only to investigate the smoothness of the solution of the differential equations and to derive stability conditions.

Instead of starting from the relation (1.10) we can start from identities which contain higher derivatives, for example

$$(1.27) \quad \begin{aligned} & y(t_{\nu+1}) - \frac{1}{2} k dy(t_{\nu+1})/dt + \frac{1}{12} k^2 d^2 y(t_{\nu+1})/dt^2 = \\ & y(t_\nu) + \frac{1}{2} k dy(t_\nu)/dt + \frac{1}{12} k^2 d^2 y(t_\nu)/dt^2 + k^5 \psi_5(t_\nu) , \end{aligned}$$

or

$$(1.28) \quad \begin{aligned} & y(t_{\nu+1}) - \frac{2}{3} k dy(t_{\nu+1})/dt + \frac{1}{6} k^2 d^2 y(t_{\nu+1})/dt^2 = \\ & y(t_\nu) + \frac{1}{3} k dy(t_\nu)/dt + k^4 \psi_4(t_\nu) , \end{aligned}$$

where

$$|\psi_5(t_\nu)| \leq \frac{1}{720} \max_{t_\nu \leq t \leq t_{\nu+1}} |d^5 y/dt^5|, \quad |\psi_4(t_\nu)| \leq \frac{1}{144} \max_{t_\nu \leq t \leq t_{\nu+1}} |d^4 y/dt^4| .$$

Observe the extremely small constant in front of $d^5 y/dt^5$. These two examples are special cases of the Pade' approximations ($p = q = 2$, and $p = 2, q = 1$ respectively)

$$(1.29) \quad \sum_{j=0}^r \frac{(-1)^j r!(r+q-j)!}{j!(r-j)!(r+q)!} k^j d^j y(t_{\nu+1})/dt^j =$$

$$\sum_{j=0}^q \frac{q!(r+q-j)!}{j!(r+q)!(q-j)!} k^j d^j y(t_{\nu})/dt^j + k^{r+q+1} \psi_{r+q+1}(t_{\nu})$$

where

$$|\psi_{r+q+1}(t_{\nu})| \leq \frac{r!q!}{(r+q)!(r+q+1)!} \max_{t_{\nu} \leq t \leq t_{\nu+1}} |d^{r+q+1} y/dt^{r+q+1}|.$$

If $y(t)$ is the solution of the differential equation (1.7) then we can rewrite the above relations as relations between $y(t_{\nu+1})$ and $y(t_{\nu})$. The coefficients then depend on A, F and their derivatives. We obtain the difference approximations by replacing $y(t_{\nu+1})$ and $y(t_{\nu})$ in these last relations by $u_{\nu+1}$ and u_{ν} , respectively, and neglecting ψ . For example, if $y(t)$ is the solution of the scalar equation (1.21) then

$$d^j y/dt^j = \lambda^j y$$

and the difference approximations corresponding to (1.27) and (1.28) are

$$(1 - \frac{1}{2} k\lambda + \frac{1}{12} k^2 \lambda^2) u_{\nu+1} = (1 + \frac{1}{2} k\lambda + \frac{1}{12} k^2 \lambda^2) u_{\nu}$$

and

$$(1 + \frac{1}{3} k\lambda) u_{\nu+1} = (1 - \frac{2}{3} k\lambda + \frac{1}{6} k^2 \lambda^2) u_{\nu}$$

respectively.

Ehle [3] has proved that the methods based on (1.29) are stable for the class η_0 if $p = q$ and strongly stable if $p = q + 1$ or $p = q + 2$. There are no new difficulties in proving theorem 1.3 also for this type of approximations.

There is a large literature on methods of this type, for example [5], [7], [9]. Sometimes they are not derived from identities between smooth functions and their derivatives. This can be dangerous.

Approximate, for example, the differential equation (1.1) by

$$(1.30) \quad (1 - ka_{11})^4 u_{\nu+1} = (1 - 3ka_{11})u_{\nu} + k(1 - ka_{11})^4 be^{dt},$$

which is strongly stable for the class η_0 . It is consistent in the usual sense, i.e. the solutions of (1.30) converge to the solution of the differential equation (1.1) as $k \rightarrow 0$. However, the convergence to the smooth part, $y_S(t)$ is not uniform for the whole class. Let $b = a_{11}$ and $y_0 + b/(a_{11} - d) = 0$. Then

$$\lim_{a_{11} \rightarrow -\infty} y = -e^{dt} \quad \text{but} \quad \lim_{a_{11} \rightarrow -\infty} |u_{\nu}| = \infty.$$

There are no difficulties to derive algebraic conditions for uniform convergence. One can develop the solutions of general systems (1.7) into asymptotic series. The same is true for the solutions of the difference approximations. Comparing these series gives algebraic conditions for uniform convergence.

2. The analytic problem.

In this section we consider the initial value problem (1.7)

$$(2.1) \quad dy/dt = A(t)y + F(t), \quad y(0) = y_0, \quad t \geq 0,$$

where $A(t)$ is negative dominant. This is no restriction, because if $A(t)$ is only essentially negative dominant then we introduce a new variable

$$\tilde{y} = e^{-\alpha t} y$$

and obtain the system

$$d\tilde{y}/dt = (A(t) - \alpha I)\tilde{y} + e^{-\alpha t} F,$$

which is negative dominant, provided α is sufficiently large. We want to estimate the solutions of (2.1) and start with

Lemma 2.1. Consider the differential equation

$$(2.2) \quad dy/dt = a_{11}(t)y + F(t), \quad a(t) = \text{Real } a_{11} \leq 0.$$

Then the following estimates hold

$$(2.3) \quad |y(t)| \leq t \max_{0 \leq \eta \leq t} |F(\eta)| + s(t)|y(0)|,$$

$$(2.4) \quad |y(t)| \leq \max_{0 \leq \eta \leq t} |F(\eta)/a(\eta)| + s(t)|y(0)| \quad \text{if } a < 0,$$

where

$$s(t) = e^{\int_0^t a(\xi) d\xi}.$$

Proof. The solution of (2.2) can be written down explicitly

$$y(t) = \int_0^t e^{\int_\eta^t a_{11}(\xi) d\xi} F(\eta) d\eta + e^{\int_0^t a_{11}(\xi) d\xi} y(0).$$

The first inequality follows from

$$\left| e^{\int_0^t a_{11}(\xi) d\xi} \right| \leq e^{\int_0^t a(\xi) d\xi} \leq 1.$$

Furthermore if $a < 0$ then

$$|y(t)| - s(t)|y(0)| \leq \left| \int_0^t e^{\int_0^\eta a_{11}(\xi) d\xi} F(\eta) d\eta \right| \leq$$

$$\left| \int_0^t a(\eta) e^{\int_0^\eta a(\xi) d\xi} |F(\eta)/a(\eta)| d\eta \right| \leq$$

$$\left| \int_0^t \frac{d}{d\eta} e^{\int_0^\eta a(\xi) d\xi} d\eta \right| \cdot \max_{0 \leq \eta \leq t} |F(\eta)/a(\eta)|,$$

which gives us (2.4).

We consider now the homogenous system

$$(2.5) \quad dv/dt = A(t)v.$$

For its solutions we have

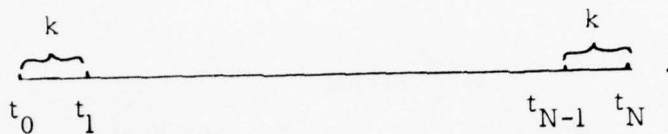
Lemma 2.2. Let t_0, t be real numbers with $0 \leq t_0 \leq t$. Assume that A is negative dominant then

$$|v(t)| \leq e^{\delta \int_{t_0}^t a(\xi) d\xi} |v(t_0)|, \quad a = \max_i \operatorname{Real} a_{ii},$$

i.e., the solution operator $S(t, t_0)$ of (2.5) satisfies the estimate

$$(2.6) \quad |S(t, t_0)| \leq e^{\delta \int_{t_0}^t a(\xi) d\xi}.$$

Proof. Let $k = (t - t_0)/N$, N natural number, and denote by $t_\nu = t_0 + \nu k$ gridpoints and by $w_\nu = w(t_\nu)$ functions defined on the grid



We approximate (2.5) by

$$(I - kA(t_{\nu+1}))w_{\nu+1} = w_\nu, \quad w_0 = v(t_0).$$

By assumption A is negative dominant. Therefore for sufficiently small k

$$(1 - \delta k a(t_{\nu+1})) |w_{\nu+1}| \leq |w_\nu|$$

i.e.

$$|w_N| \leq \prod_{\nu=0}^{N-1} (1 - \delta k a(t_{\nu+1}))^{-1} |w_0|.$$

We know that, as $k \rightarrow 0$,

$$w_N \rightarrow v(t), \quad \prod_{\nu=0}^{N-1} (1 - \delta k a(t_{\nu+1}))^{-1} \rightarrow e^{\delta \int_{t_0}^t a(\xi) d\xi}.$$

Therefore the estimate (2.6) follows.

For the system (2.1) we have

Lemma 2.3. Assume that A is negative dominant. Then the solution of (2.1) satisfies the estimates

$$(2.7) \quad |y(t)| \leq t \max_{0 \leq \eta \leq t} |F(\eta)| + s(t) |y(0)| ,$$

$$(2.8) \quad |y(t)| \leq 2\delta^{-1} \max_{0 \leq \eta \leq t} |(\Lambda(\eta) + \Lambda^*(\eta))^{-1} F(\eta)| + s(t) |y(0)| ,$$

where

$$s(t) = e^{\delta \int_0^t a(\xi) d\xi} , \quad a = \max_i \text{Real } a_{ii} ,$$

and

$$\Lambda = \begin{pmatrix} a_{11} & 0 & . & . & . & . & 0 \\ 0 & a_{22} & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & . & . & . & . & . & a_{nn} \end{pmatrix} .$$

Proof. The first estimate follows from lemma 2.2 and the representation

$$y(t) = \int_0^t S(t, \eta) F(\eta) d\eta + S(t, 0) y(0) .$$

We are now going to prove the second estimate. By lemma 2.2 we can assume that $y(0) = 0$. Let t be fixed and denote by M the space of all continuous functions $g(t)$ with $g(0) = 0$. The system

$$\mathfrak{L}_0 y \equiv dy/dx - \Lambda y = F$$

has a unique solution in M and, by lemma 2.1

$$(2.9) \quad \|x_0^{-1}F\| \leq 2\|(\Lambda + \Lambda^*)^{-1}F\|, \quad \|g\| = \max_{0 \leq \xi \leq t} |g(\xi)|.$$

Let x_1 denote the operator

$$x_1 y = (A - \Lambda)y, \quad y \in M.$$

Then we can write the differential equation (2.1) in the form

$$(I - x_0^{-1}x_1)y = x_0^{-1}F.$$

Definition 1.1 and (2.9) imply

$$\|x_0^{-1}x_1 y\| \leq 2\|(\Lambda + \Lambda^*)^{-1}x_1 y\| \leq (1 - \delta)\|y\|,$$

and the estimate (2.8) follows from (2.9).

The assumption that A is negative dominant implies that

$$(2.10) \quad |(\Lambda + \Lambda^*)^{-1}A| \leq |(\Lambda + \Lambda^*)^{-1}\Lambda| + |(\Lambda + \Lambda^*)^{-1}(A - \Lambda)| \leq \frac{1}{2}(\rho + 1 - \delta).$$

Thus (2.8) can also be expressed as

$$(2.11) \quad |y(t)| \leq \frac{\rho + 1 - \delta}{\delta} \max_{0 \leq \eta \leq t} |A^{-1}(\eta)F(\eta)| + s(t)|y(0)|.$$

We shall now prove theorem 1.2. We use the notation $F^{[\nu]} = d^\nu F/dt^\nu$ and start with

Lemma 2.4. Assume that A is negative dominant. $A^{-1}A^{[\nu]}$, $A^{-1}F^{[\nu]}$, $\nu = 0, 1, 2, \dots, p$; can be estimated by the functions (1.8).

Proof. Denote by Λ the diagonal of A . Then

$$|A^{-1}A^{[\nu]}| \leq (I + \Lambda^{-1}(A - \Lambda))^{-1} |\Lambda^{-1}A^{[\nu]}|.$$

By assumption $|a_{ii}| \geq 1$ and $|(I + \Lambda^{-1}(A - \Lambda))^{-1}| < \delta^{-1}$. Thus $A^{-1}A^{[\nu]}$ can be estimated by $\Lambda^{-1}A^{[\nu]}$, which is composed of terms $a_{ii}^{-1}a_{ij}^{[\nu]}$. The same procedure works for $A^{-1}F^{[\nu]}$ and the lemma is proved.

Lemma 2.5. Assume that A is negative dominant and define $C(t)$ by $A(t) = A(0)C(t)$. Then $C^{[\nu]}$, $(C^{-1})^{[\nu]}$, $\nu = 0, 1, 2, \dots, p$; can be estimated by $A^{-1}A^{[\nu]}$, $\nu = 0, 1, 2, \dots, p$.

Proof. Let $B = A^{-1}dA/dt$. Then A^* is the solution of

$$dA^*/dt = B^* A^*.$$

$C^*(t)$ is the solution operator of this differential equation and therefore the lemma follows from well known estimates.

Lemma 2.6. Assume that A is negative dominant. The solution $y(t)$ of (2.1) and its first p derivatives can be estimated by

$$(2.12) \quad y^{[\nu]}(0), A^{-1}A^{[\nu]}, A^{-1}F^{[\nu]}, \quad \nu = 0, 1, 2, \dots, p.$$

Proof. Differentiating (2.1) gives us

$$(2.13) \quad dy^{[\nu]}/dt = Ay^{[\nu]} + \sum_{\mu=0}^{\nu-1} \binom{\nu}{\mu} A^{[\nu-\mu]} y^{[\mu]} + F^{[\nu]}.$$

Therefore the lemma follows by induction from (2.11).

A simple consequence of lemma 2.6 is

Lemma 2.7. Assume that A is negative dominant and that

$$(2.14) \quad y(0) = 0, F^{[\nu]}(0) = 0, \quad \nu = 0, 1, 2, \dots, p-1.$$

Then the solution $y(t)$ of (2.1) and its first p derivatives can be estimated by

$$(2.15) \quad A^{-1}A^{[\nu]}, A^{-1}F^{[\nu]}, \quad \nu = 0, 1, 2, \dots, p.$$

Proof. (2.13) implies $y^{[\nu]}(0) = 0$ for $\nu = 0, 1, 2, \dots, p$. Therefore the lemma follows from lemma 2.6.

We can now prove the first part of theorem 1.2 by reducing the general case to the special case of lemma 2.7. Consider the system

$$(2.16) \quad dw/dt = \tilde{A}w + \tilde{F},$$

where

$$\tilde{A} = \sum_{\nu=0}^{p-1} \frac{t^\nu}{\nu!} A^{[\nu]}(0), \quad \tilde{F} = \sum_{\nu=0}^{p-1} \frac{t^\nu}{\nu!} F^{[\nu]}(0),$$

in a neighbourhood of $t = 0$. We seek a solution of the form

$$(2.17) \quad w(t) = \sum_{\nu=0}^{p-1} \frac{t^\nu}{\nu!} w_\nu + \varphi(t), \quad \varphi(0) = 0.$$

Introducing (2.17) into (2.16) gives us

$$(2.18) \quad d\varphi/dt = \tilde{A}\varphi + f(t), \quad \varphi(0) = 0,$$

with

$$f(t) = \sum_{\nu=0}^{p-1} \frac{t^\nu}{\nu!} \left(\sum_{\mu=0}^{\nu} A^{[\mu]}(0) \frac{\nu! w_{\nu-\mu}}{\mu!(\nu-\mu)!} - w_{\nu+1} + F^{[\nu]}(0) \right) + t^p f_1(t),$$

where $w_p = 0$ and $f_1(t)$ is a polynomial in t with coefficients depending linearly on $A^{[\nu]}(0)w_\mu$. Choose now the w_ν such that

$$(2.19) \quad \sum_{\mu=0}^{\nu} A^{[\mu]}(0) \frac{\nu! w_{\nu-\mu}}{\mu!(\nu-\mu)!} - w_{\nu+1} + F^{[\nu]}(0) = 0, \quad \nu = 0, 1, 2, \dots, p-1.$$

We shall show that this is always possible. Let

$$L_\nu = - \sum_{\mu=1}^{\nu} A^{-1}(0)A^{[\mu]} \frac{\nu! w_{\nu-\mu}}{\mu!(\nu-\mu)!}.$$

Then (2.19) can be written in the form

$$(2.20) \quad w_\nu - A^{-1}(0)w_{\nu+1} = L_\nu - A^{-1}(0)F^{[\nu]}(0), \quad \nu = 0, 1, 2, \dots, p-1; L_0 = 0.$$

Observe that L_ν depends only on w_μ with $\mu < \nu$ and that $w_p = 0$.

Therefore, if we neglect the term $A^{-1}(0)w_{\nu+1}$ then the w_ν are uniquely determined by (2.20) and can be estimated by $A^{-1}(0)A^{[\nu]}(0)$, $A^{-1}(0)F^{[\nu]}(0)$.

The same is true if $|A^{-1}(0)|$ is sufficiently small. If $A^{-1}(0)$ is not small, then the determinant D of the linear system (2.20) can vanish.

However, by introducing a new variable $\tilde{y} = \exp(\alpha t)y$ into (2.1), we can always choose α such that $D \neq 0$. Observe that D does not vanish identically.

In a neighbourhood of $t = 0$ the system (2.18) satisfies the conditions of lemma 2.7. Furthermore the derivatives $\tilde{A}^{-1}\tilde{A}^{[\nu]}$, $\tilde{A}^{-1}\tilde{F}^{[\nu]}$ can be estimated by $A^{-1}A^{[\nu]}$, $A^{-1}F^{[\nu]}$. Therefore, also $\varphi(t)$ can be estimated by these functions. This is also true for the solution of (2.16) if we choose

$$(2.21) \quad w(0) = w_0, \quad \text{i.e.,} \quad \varphi(0) = 0$$

as the initial value.

Let $g(t) \in C^\infty$ be a "cut-off" function, i.e. there are constants α_1, α_2 with $0 < \alpha_1 < \alpha_2 < \infty$ such that

$$g(t) = \begin{cases} 1 & \text{for } 0 \leq t \leq \alpha_1, \\ 0 & \text{for } t \geq \alpha_2. \end{cases}$$

Let w be the solution of (2.16). Then $\tilde{w} = gw$ is the solution of

$$(2.22) \quad d\tilde{w}/dt = \tilde{A}\tilde{w} + \tilde{F}, \quad \tilde{F} = (dg/dt)w + g\tilde{F}, \quad \tilde{w}(0) = w_0.$$

If we choose the α_j , $j = 1, 2$; sufficiently small then \tilde{w} has the same properties as y in lemma 2.7. Subtract (2.22) from (2.1). Then

$u = y - \tilde{w}$ is the solution of

$$(2.23) \quad du/dt = Au + (A - \tilde{A})\tilde{w} + F - \tilde{F}, \quad u(0) = y(0) - w_0.$$

u can be written as $u = u_1 + v$ where

$$dv/dt = Av, \quad v(0) = y(0) - w_0,$$

and u_1 is the solution of the differential equation (2.23) with homogenous initial value. For u_1 the conditions of lemma 2.7 are fulfilled and the first part of the lemma is proved.

To prove the second part of theorem 1.2 we need another version of lemma 2.7.

Lemma 2.7a. Assume that A is negative dominant and that

$$y(0) = 0, \quad F^{[\nu]}(0) = 0, \quad \nu = 0, 1, 2, \dots, p-2.$$

Then the solution of (2.1) and its first p derivatives can be estimated by

$$A^{-1}A^{[\nu]}, \quad A^{-1}F^{[\nu]}, \quad \nu = 0, 1, 2, \dots, p; \quad \text{and} \quad F^{[p-1]}(0).$$

Proof. (2.13) implies $y^{[\nu]}(0) = 0$, $\nu = 0, 1, 2, \dots, p$; and $y^{[p]}(0) = F^{[p-1]}(0)$. Therefore the lemma follows from lemma 2.6.

We need also

Lemma 2.8. Assume that $A(t)$ is negative dominant. Let $v(t)$ be a solution of (2.5). Then

$$(t^\nu v)^{[\mu]}, t^\nu v^{[\mu]}, \nu \geq \mu, \nu, \mu = 0, 1, 2, \dots, p;$$

can be estimated by

$$A^{-1}A^{[\tau]}, \tau = 0, 1, 2, \dots, p; \text{ and } v(0).$$

Proof. It is clear, that we need to prove the theorem only for $t^\nu v^{[\mu]}$.

Let $w = t^\nu v$. Then w is the solution of

$$dw/dt = Aw + \nu t^{\nu-1}v, w(0) = 0.$$

We have to prove that $w^{[\mu]}, \mu = 0, 1, 2, \dots, \nu;$ can be estimated by $A^{-1}A^{[\tau]}, \tau = 0, 1, 2, \dots, p$. For $p = 1$ this follows directly from lemma 2.6 because $A^{-1}dv/dt = v$. The general case follows by induction using lemma 2.6, and the observation that $v^{[\mu]}$ can be expressed by derivatives of lower order using the differential equation (2.5).

The last lemma shows, that outside a "small" neighbourhood of $t = 0$, the solution $v(t)$ of (2.5) and its derivatives can be estimated by $A^{-1}A^{[\nu]}$. We shall now improve these estimates. We shall show, that if $A^{-1}dA/dt$ is of moderate size, then the rapidly changing part of $v(t)$ is to a first approximation the solution of the system

$$(2.24) \quad dv_1/dt = A(0)v_1, v_1(0) = v(0),$$

with constant coefficients.

Theorem 2.2. Assume that A is negative dominant. Then the solution of (2.5) can be written as

$$(2.25) \quad v = v_1 + e_1$$

where v_1 is the solution of (2.24) and e_1 is the solution of

$$(2.26) \quad de_1/dt = A(t)e_1 + F(t), \quad F(t) = (A(t) - A(0))v_1(t), \quad e_1(0) = 0.$$

Furthermore e_1 and de_1/dt can be estimated by $A^{-1}dA/dt$.

Proof. Introducing (2.25) into (2.5) gives us (2.26). By lemma 2.5

$$A^{-1}(t)F(t) = \frac{1 - C^{-1}(t)}{t} A^{-1}(0)tv_1.$$

Lemma 2.8, applied to the equation (2.24), and lemma 2.5 show, that the theorem follows from lemma 2.6 with $p = 1$.

In the same way one can prove.

Theorem 2.3. Assume that A is negative dominant. Then the solution of (2.5) can be written as

$$(2.27) \quad v(t) = \sum_{\mu=1}^p v_{\mu}(t) + e_p(t)$$

where $v_1(t)$ is the solution of (2.24), the $v_{\mu}(t)$, $\mu > 1$ are the solutions of

$$(2.28) \quad dv_{\mu+1}/dt = A(0)v_{\mu+1}(t) + (A(t) - A(0))v_{\mu}, \quad v_{\mu+1}(0) = 0, \\ \mu = 1, 2, \dots, p-1;$$

and e_p satisfies the equation

$$(2.29) \quad de_p/dt = A(t)e_p + (A(t) - A(0))v_p(t), \quad e_p(0) = 0.$$

Furthermore

$$e_p^{[\nu]}, \quad \nu = 0, 1, 2, \dots, p, \quad v_{\mu}^{[\nu]}, \quad \nu = 0, 1, 2, \dots, \mu; \quad \mu = 1, 2, \dots, p-1;$$

can be estimated by

$$v(0) \text{ and } A^{-1}A^{[\mu]}, \mu = 1, 2, \dots, p.$$

Finally, in the same way as before, we can replace $A(t)$ by

$\sum_{\nu=0}^{p-1} \frac{t^\nu}{\nu!} A^{[\nu]}(0)$ without destroying the desired estimates. The resulting equations can be solved explicitly and consist of terms of the form $t^n e^{\lambda t} a$ where λ denote the eigenvalues of $A(0)$. These terms generate at most boundary layers. We have thus proved theorem 1.2.

3. A normal form for the differential equations.

In the first section we discussed examples which show how important it is, that the system (1.7) is essentially negative dominant and that the functions (1.8) are of moderate size. One can also give another argument. For simplicity we consider only the homogenous system

$$(3.1) \quad dy/dt = A(t)y$$

where $A(t) \in C^p$, $p \geq 1$ but A and its derivatives need not be of moderate size. Assume that we want to solve this system by a standard difference method. Then we discretize (3.1) and use point values $A(t_0)$. Therefore it is reasonable to demand, that the equations

$$dw/dt = A_0 w, \quad A_0 = A(t_0)$$

with constant coefficients, locally describe the system (3.1). For the numerical technique to work, it is essential, that the solutions of (3.1) do not grow too fast. Thus, we demand that there are constants K_0, c of moderate size such that

$$(3.2a) \quad |e^{A_0 t}| \leq K_0 e^{ct}$$

for all fixed values t_0 . (3.2a) implies that the eigenvalues λ of A_0 satisfy the condition

$$\text{Real } \lambda \leq c.$$

Another demand is that the solutions do not oscillate too rapidly. This leads to the condition

$$(3.2b) \quad |\text{Im } \lambda| \leq \rho |\text{Real } \lambda| + c, \quad \rho, c \text{ constants of moderate size.}$$

We shall show that the conditions (3.2a) and (3.2b) naturally lead to matrices which are essentially negative dominant. We have

Theorem 3.1. Let $\mathfrak{F}_1(K_0, c, \rho)$ denote the family of all $n \times n$ matrices for which (3.2a) and (3.2b) hold, and let $\mathfrak{F}_2(c, \rho, \delta)$ denote the family of $n \times n$ matrices which are essentially negative dominant. For any fixed K_0, δ there is a universal constant K_1 such that for every matrix $A \in \mathfrak{F}_1(K_0, c, \rho)$ there is a nonsingular transformation S with

$$|S| |S^{-1}| \leq K_1 \text{ and } S^{-1}AS \in \mathfrak{F}_2(c, \rho, \delta).$$

Proof. In [6] we have shown that there are universal constants K_{11}, K_{12} such that for every $A \in \mathfrak{F}_1$ there is a nonsingular transformation T with

$$|T| |T^{-1}| \leq K_{11}$$

and

$$T^{-1}(A - cI)T = \begin{pmatrix} b_{11} & b_{12} & \cdot & \cdot & \cdots & b_{1n} \\ 0 & b_{22} & b_{23} & \cdots & b_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & b_{nn} \end{pmatrix},$$

where

$$\sum_{\substack{j=1 \\ j \neq i}}^n |b_{ij}| \leq K_{12} |\text{Real } b_{ii}|, \text{ Real } b_{ii} \leq 0, \quad i = 1, 2, \dots, n.$$

By (3.2b)

$$|\text{Im } b_{ii}| \leq \rho |\text{Real } b_{ii}| + c, \quad i = 1, 2, \dots, n.$$

Therefore we can choose a diagonal scaling

$$D = \begin{pmatrix} d_1 & 0 & . & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & d_n \end{pmatrix}, \quad d_i > 0$$

such that $S^{-1}AS$, $S = TD$ is essentially negative dominant, i.e. belongs to \mathfrak{F}_2 . Here $|D||D^{-1}|$ only depends on δ , K_{12} .

Now assume that $A(t) \in \mathfrak{F}_1$ for every fixed t and consider the system (3.1) in a neighbourhood of a point t_0 . Let S_0 denote the transformation of the last theorem. Introducing a new variable $\tilde{y} = S_0^{-1}y$ we obtain

$$(3.3) \quad d\tilde{y}/dt = B(t)\tilde{y}, \quad B(t) = S_0^{-1}A(t)S_0,$$

where $B(t_0)$ is essentially negative dominant. If $A(t_0)$ shall represent $A(t)$ in a whole neighbourhood of t_0 then $B(t)$ must be essentially negative dominant in a neighbourhood of t_0 provided we change δ to $\frac{1}{2}\delta$ and c to $2c$. Thus, we can divide our interval of integration into subintervals $t_i \leq t \leq t_{i+1}$ and in every subinterval transform $A(t)$ into an essentially negative dominant matrix by a transformation S_i for which $|S_i||S_i^{-1}|$ is uniformly bounded. A more precise description is given in

Theorem 3.2. Consider the system (3.1) and assume that there are constants K, c, ρ such that the matrices $A(t) \in \mathfrak{F}_1(K, c, \rho)$ for every fixed t . Assume also that there is a constant K_2 such that for all fixed t

$$|\tilde{A}^{-1}(t)d\tilde{A}/dt| \leq K_2, \quad \tilde{A} = A - (c+1)I.$$

Then one can divide the interval of integration into subintervals

$$t_i \leq t \leq t_{i+1}, \quad t_{i+1} - t_i \geq \eta > 0, \quad \eta = \text{const.},$$

and in every subinterval there is a transformation S_i with $|S_i| |S_i^{-1}|$ uniformly bounded and $S_i^{-1} A(t) S_i \in \mathfrak{F}_2(C+1, 2\rho, \delta)$. Furthermore δ can be chosen arbitrarily.

Proof. Let t_0 be a fixed point. Without restriction we can assume that $A(t_0)$ is essentially negative dominant. Otherwise we would use the transformation of theorem 3.1 to obtain the system (3.3). Proceeding as in lemma 2.5

$$\tilde{A}(t - t_0) = \tilde{A}(t_0) E(t - t_0)$$

with

$$E(0) = I \quad \text{and} \quad |dE/dt| = |\tilde{A}^{-1} d\tilde{A}/dt| \leq K_2$$

and the theorem follows.

If the system does not satisfy the above conditions then using a numerical procedure directly can be very dangerous. However, we shall show that we can obtain the normal form by combining the transformation of the dependent variable with a local stretching of the independent variable.

We divide the interval of integration into subintervals $t_i \leq t \leq t_{i+1}$ and construct in every subinterval a nonsingular transformation S_i and a stretching

$$t - t_i = \alpha_i(\tilde{t} - i), \quad \alpha_i = t_{i+1} - t_i, \quad i \leq \tilde{t} \leq i+1,$$

such that the transformed systems

$$(3.4) \quad dv/d\tilde{t} = \alpha_i S_i^{-1} A(\tilde{t}) S_i v$$

are essentially negative dominant and the functions (1.8) are of moderate size. Then the solutions of (3.4) are smooth in the interior of every subinterval $i \leq \tilde{t} \leq i+1$, but boundary layers could appear at $\tilde{t} = 1, 2, \dots$. By lemma 2.6 we can avoid these by demanding that the stretching constants α_i do not increase too fast and that the transformations S_i do not change too fast, i.e. there are constants $d > 1$ and $K_0 > 1$ of moderate size, such that

$$(3.5) \quad \alpha_{i+1}/\alpha_i \leq d, \max(|S_i^{-1} S_{i+1}|, |S_{i+1}^{-1} S_i|) \leq K_0.$$

We shall now describe the details and start with the scalar equation

$$(3.6) \quad \varepsilon dy/dt = -t^p y, y(0) = y_0, t \geq 0,$$

where p is a natural number and $\varepsilon > 0$ a small constant. Its solution is given by $y = \exp(t^{p+1}/\varepsilon(p+1)) y_0$, i.e. it is a function of $t/\varepsilon^{1/(p+1)}$.

In a neighbourhood of $t = 0$ we introduce into (3.6) a new variable \tilde{t} by $t = \alpha \tilde{t}$, $\alpha = \text{const.} > 0$ and obtain

$$(3.7) \quad dy/d\tilde{t} = -\frac{\alpha^{p+1} \tilde{t}^p}{\varepsilon} y = a_{11}(\tilde{t}) y.$$

We determine the largest α such that

$$(3.8) \quad \max_{0 \leq \tilde{t} \leq 1} (\min(1, |a_{11}^{-1}(\tilde{t})|) |da_{11}(\tilde{t})/d\tilde{t}|) \leq K.$$

Here $K \geq 1$ is a threshold constant of moderate size. For simplicity we assume that $K = p$ and obtain

$$p\alpha^{p+1} = K\varepsilon, \quad \text{i.e. } \alpha = \varepsilon^{1/(p+1)}.$$

Thus we have determined the stretched variable \tilde{t} for $0 \leq t \leq \varepsilon^{1/(p+1)} = t_1$ (3.8) guarantees that the solution of the transformed differential equation (3.7) is smooth for $0 \leq \tilde{t} \leq 1$. (In general one has to include more derivatives in the expression (3.8). However, for the equation (3.6) the inequality (3.8) insures that $a_{11}(\tilde{t})$ and all its derivatives are of moderate size. This is quite common.)

We determine now the stretched variable in a neighbourhood of $t = t_1$. Let $t = t_1 + \alpha(\tilde{t} - 1)$, then the differential equation becomes

$$dy/d\tilde{t} = \frac{(t_1 + (\tilde{t} - 1))^p}{\varepsilon} y = a_{11}(\tilde{t})y.$$

The obvious modification of (3.8) gives us $\alpha = t_1$ and we have determined the stretched variables for $0 \leq t \leq t_2 = 2t_1$. This process can be continued. After n steps we have obtained the stretched variables for $0 \leq t \leq t_n = 2^{n-1}t_1$. It is clear, that the interval length corresponds exactly to the behavior of the solution of (3.6).

If the given system of differential equations (3.1) is essentially negative dominant, then we can use the corresponding procedure. In the simplest case the condition

$$(3.9) \quad \max_{i \leq \tilde{t} \leq i+1} (\min(1, |a_{ii}^{-1}(\tilde{t})|) |da_{ij}(\tilde{t})/d\tilde{t}|) \leq K, \quad i, j = 1, 2, \dots, n;$$

determines a stretching

$$(3.10) \quad t - t_i = \alpha_i(\tilde{t} - \tilde{t}_i), \quad \tilde{t}_{i+1} - \tilde{t}_i = 1.$$

Applying (3.9) to (3.1) gives us

$$(3.11) \quad dy/d\tilde{t} = \alpha_i A(\tilde{t})y.$$

If $\alpha_i \leq 1$ then also (3.11) is essentially negative dominant. If $\alpha_i > 1$ then this need not be so and we have to limit α_i to a value $\alpha_i \geq 1$ such that (3.11) satisfies (1.3a), (1.4a) and (1.4b).

If the system (3.1) is not essentially negative dominant, then we have to transform it to that form. This can be done in the following way. For $t = 0$ we use the Q - R method to construct a unitary matrix $U(0)$ which transforms $A(0)$ to upper triangular form

$$U^*(0)A(0)U(0) = \begin{pmatrix} \kappa_1 & b_{12} & . & \cdots & b_{1n} \\ 0 & \kappa_2 & b_{23} & \cdots & b_{2n} \\ . & . & . & . & . \\ 0 & . & . & 0 & \kappa_n \end{pmatrix} = B_1(0),$$

where $|\kappa_1| \geq |\kappa_2| \geq \cdots \geq |\kappa_n|$. The next step is to check whether the conditions (1.3a) and (1.4a) are satisfied. If not, we apply a stretching $t = \beta\hat{t}$, $0 \leq \hat{t} \leq 1$, such that the eigenvalues of $B_2(0) = \beta B_1(0)$ satisfy these conditions. (This procedure is not very satisfactory and shall be improved in another paper.)

The last step consists of a diagonal scaling

$$D_0 = \begin{pmatrix} d_1 & 0 & . & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & d_n \end{pmatrix},$$

such that

$$(3.12) \quad D_0^{-1} B_2(0) D_0 = B_3(0)$$

satisfies the condition (1.4b).

Unfortunately $|D| |D|^{-1}$ is not always of moderate size. In that case we proceed in the following way. Let $\tau > 1$ be a constant. We shall divide the eigenvalues β_{κ_i} of $B_2(0)$ into classes N_j . The class N_1 is defined by

- 1) $\beta_{\kappa_i} \in N_1$ if $|\beta_{\kappa_i}| \leq c$,
- 2) $\beta_{\kappa_n} \in N_1$,
- 3) $\beta_{\kappa_i} \in N_1$ if $\beta_{\kappa_{i+1}} \in N_1$ and $|\kappa_i|/|\kappa_{i+1}| \leq \tau$.

If N_1 does not contain all eigenvalues then there is one eigenvalue with $\beta_{\kappa_{p+1}} \in N_1$ but $\beta_{\kappa_p} \notin N_1$. Then the class N_2 is defined by the last two rules with β_{κ_n} replaced by β_{κ_p} . The other classes are constructed correspondingly.

Now we construct a transformation H such that

$$H^{-1} B_2(0) H = \begin{pmatrix} B_{rr} & 0 & . & \dots & 0 \\ 0 & B_{r-1 \ r-1} & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & \dots & B_{11} \end{pmatrix}$$

where every B_{jj} is of the same form as $B_2(0)$ and its eigenvalues consist of the class N_j . Then we apply a diagonal scaling

$$D_1 = \begin{pmatrix} D_1 & 0 & . & \dots & 0 \\ 0 & D_2 & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & D_{rr} \end{pmatrix}$$

such that

$$(3.13) \quad B_3(0) = D_1^{-1} H^{-1} B_2(0) D_1 H$$

is essentially negative dominant. If $|D_1 H| |(D_1 H)^{-1}| < |D_0| |D_0^{-1}|$ we use (3.13) otherwise we return to (3.12).

In this way we have constructed a transformation $v = S_0^{-1} y$ and a stretching $t = \beta \hat{t}$, $0 \leq \hat{t} \leq 1$, such that the system

$$dv/d\hat{t} = \beta S_0^{-1} A(\hat{t}) S_0 v$$

has the property that $\beta S_0^{-1} A(0) S_0$ is essentially negative dominant. Then there is an interval $0 \leq \hat{t} \leq \hat{t}_1$ in which the same is true for $\beta S_0^{-1} A(\hat{t}) S_0$ if we replace c by $2c$ and δ by $\frac{1}{2} \delta$. Let

$$\tilde{t} = \begin{cases} \hat{t} & \text{if } \hat{t}_1 \geq 1 \\ \hat{t}/\hat{t}_1 & \text{if } \hat{t}_1 < 1 \end{cases}$$

then the system

$$(3.14) \quad dv/d\tilde{t} = \alpha_0 S_0^{-1} A(\tilde{t}) S_0 v$$

is essentially negative dominant for $0 \leq \tilde{t} \leq 1$.

At $\hat{t} = 1$ we modify this process somewhat. By (3.5) the stretching shall not increase too fast. Also, in many cases S_0 is a good approximation for S_1 . Therefore we perform at $t = t_1 = \alpha_0$ a preliminary scaling $t - t_1 = d\alpha_0(\hat{t} - 1)$, $1 \leq \hat{t} \leq 2$, and the transformation $u = S_0^{-1}y$ to obtain

$$(3.15) \quad du/d\hat{t} = d\alpha_0 S_0^{-1} A(\hat{t}) S_0 u.$$

If (3.15) is essentially negative dominant then we use (3.15). Otherwise we use the process as described earlier to construct a matrix $T(t_1)$ such that

$$(3.16) \quad \beta d\alpha_0 T^{-1}(t_1) S_0^{-1} A(t_1) S_0 T(t_1)$$

is essentially negative dominant. If $|T||T^{-1}|$ is of moderate size then we proceed as earlier. Otherwise we return to (3.14) and replace α_0 and t_1 by $\alpha_0/2$ and $t_1/2$ respectively and restart the process.

Eventually the procedure will stop because in the worst situation $d(\alpha_0/2^p) S_0^{-1} A(t_1(2^p)) S_0$ will be of moderate size. Then we can choose $T = I$, i.e. $S_0 = S_1$ and proceed as earlier. In this way we determine t_2 and the following t_j and construct a system which is piecewise negative dominant. The new system is treated as stated earlier (see (3.9)).

We shall now illustrate the technique for a simple example. Consider the differential equation

$$\epsilon y'' + (a(t)y)' + b(t)y = 0, \quad t \geq 0$$

with initial conditions

$$y(0) = y_0, \quad y'(0) = y_1.$$

Here $\epsilon > 0$ is a small constant and a, b with $a \geq 0, b \leq b_0 < 0$ are smooth functions of moderate size. We write the above equation as a first order system by introducing new variables $y^{(1)} = y, dy^{(2)}/dt = b(t)y$ and obtain

$$(3.17) \quad \frac{dy}{dt} = \frac{d}{dt} \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} = \begin{pmatrix} -\frac{a(t)}{\epsilon} & -\frac{1}{\epsilon} \\ b(t) & 0 \end{pmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} = A(t)y.$$

If $a(t) > 1$ and $|b(t)| < c$ then the system is essentially negative dominant and nothing needs to be done. Otherwise we transform the matrix A for $t = 0$ to upper triangular form employing a unitary transformation $U = U(0)$. The eigenvalues λ of $A(0)$ are the solutions of the characteristic equation

$$\epsilon \lambda^2 + a(0)\lambda + b(0) = 0,$$

i.e.

$$\lambda_1 = -\frac{a(0)}{2\epsilon} - \text{sign } a \cdot \sqrt{-\frac{b(0)}{\epsilon} + \frac{a^2(0)}{4\epsilon^2}},$$

$$\lambda_2 = -\frac{a(0)}{2\epsilon} + \text{sign } a \cdot \sqrt{-\frac{b(0)}{\epsilon} + \frac{a^2(0)}{4\epsilon^2}}.$$

Observe that for $|a| \gg \sqrt{\epsilon}$ the two eigenvalues have the form

$$\lambda_1 \sim -\frac{a(0)}{\epsilon}, \quad \lambda_2 = \frac{b(0)}{\epsilon \lambda_1} \sim -\frac{b(0)}{a(0)}.$$

Furthermore $b < 0$ implies that λ_1, λ_2 are real and

$$|\lambda_1(0)| \geq |b(0)/\epsilon|^{1/2}.$$

The eigenvector corresponding to λ_1 is given by

$$e_1 = \tau \begin{pmatrix} 1 \\ b(0)/\lambda_1(0) \end{pmatrix}, \quad \tau = (1 + (b(0)/\lambda_1(0))^2)^{-1/2}$$

and the desired unitary transformation has the form

$$U(0) = \tau \begin{pmatrix} 1 & -\frac{b(0)}{\lambda_1(0)} \\ +\frac{b(0)}{\lambda_1(0)} & 1 \end{pmatrix}.$$

Then

$$B_1(t) = U^*(0)A(t)U(0) \approx \begin{pmatrix} \lambda_1(0)g(t) - \frac{1}{\epsilon} & \\ \frac{d(t)}{\epsilon\lambda_1(0)} & \frac{b(0)}{\epsilon\lambda_1(0)} \end{pmatrix}$$

with $d(t) = b(0)(a(t) - a(0))$, and $g(t) = 1 + \frac{a(0) - a(t)}{\epsilon\lambda_1(0)}$. Thus $d(0) = 0$, $g(0) = 1$

and $B_1(t)$ is upper triangular for $t = 0$. The condition (1.3a) is always

satisfied. However, (1.4a) does not need to hold, because $\lambda_2 > 0$

and $\lambda_2 = 0(1/\sqrt{\epsilon})$ when $a(t) = 0$. To satisfy condition (1.4a) we

perform a stretching $t = \beta\hat{t}$, where

$$(3.18) \quad \beta = \begin{cases} 1 & \text{if } \frac{b(0)}{\epsilon\lambda_1(0)} \leq \frac{1}{2}c, \\ \frac{c\epsilon\lambda_1(0)}{2b(0)} & \text{if } \frac{b(0)}{\epsilon\lambda_1(0)} > \frac{1}{2}c. \end{cases}$$

Then

$$B_2(\hat{t}) = \beta B_1(\hat{t}) \approx \begin{pmatrix} \beta \lambda_1(0) g(\beta \hat{t}) & -\frac{\beta}{\varepsilon} \\ \frac{\beta d(\beta \hat{t})}{\varepsilon \lambda_1(0)} & \frac{\beta b(0)}{\varepsilon \lambda_1(0)} \end{pmatrix}.$$

The next step is to apply a diagonal scaling

$$D(0) = \begin{pmatrix} \frac{2}{\varepsilon \lambda_1(0)} & 0 \\ 0 & 1 \end{pmatrix}.$$

Let $v = D^{-1} U^* y$. Then (3.17) becomes

$$(3.19) \quad \frac{dv}{dt} \approx \begin{pmatrix} \beta \lambda_1(0) g(\beta \hat{t}) & -\frac{\beta}{2} \lambda_1(0) \\ \frac{2\beta d(\beta \hat{t})}{\varepsilon^2 \lambda_1^2(0)} & \frac{\beta b(0)}{\varepsilon \lambda_1(0)} \end{pmatrix} v.$$

Denote by

$$(3.20) \quad \hat{t}_1 \leq \frac{c\varepsilon \lambda_1(0)}{2\beta b(0)} = \begin{cases} \frac{c\varepsilon \lambda_1(0)}{2b(0)} & \text{if } \frac{b(0)}{\varepsilon \lambda_1(0)} \leq \frac{1}{2}c, \\ 1 & \text{if } \frac{b(0)}{\varepsilon \lambda_1(0)} > \frac{1}{2}c, \end{cases}$$

the largest value such that for all \hat{t} with $0 \leq \hat{t} \leq \hat{t}_1$

$$(3.21) \quad \frac{a(0) - a(\beta \hat{t})}{\varepsilon \lambda_1(0)} \geq -\frac{1}{4}, \quad \left| \hat{t}_1 \frac{2\beta d(\beta \hat{t})}{\varepsilon^2 \lambda_1^2(0)} \right| \leq \frac{3}{2}c,$$

and introduce a new variable \tilde{t} by $\hat{t} = \hat{t}_1 \tilde{t}$. Then (3.19) becomes

$$(3.22) \quad \frac{dv}{d\tilde{t}} \approx \hat{t}_1 \begin{pmatrix} \beta \lambda_1(0) g(\beta \hat{t}_1 \tilde{t}) - \frac{\beta}{2} \lambda_1(0) \\ \frac{2\beta d(\beta \hat{t}_1 \tilde{t})}{\varepsilon^2 \lambda_1^2(0)} \quad \frac{\beta b(0)}{\varepsilon \lambda_1(0)} \end{pmatrix}^v,$$

and the last system is essentially negative dominant for $0 \leq \tilde{t} \leq 1$. The conditions (3.20) and (3.21) are satisfied if we choose

$$(3.23) \quad t_1 = \beta \hat{t}_1 = c_1 \varepsilon |\lambda_1(0)|.$$

Here c_1 is a constant which depends on c , b and da/dt but not on ε . Also, the conditions (3.9) are essentially satisfied and no further stretching is necessary. Thus, the interval length $0 \leq t \leq t_1$ is given by $t_1 = \text{const. } \varepsilon |\lambda_1(0)|$. In general we get

$$t_{i+1} - t_i = \text{const. } \varepsilon |\lambda(t_i)|.$$

There are two different situations

1) $a(t) \geq a_0 > 0$ everywhere. Then $\varepsilon |\lambda_1| \approx |a/b|$ and the interval length never becomes very small.

2) $a(t) = 0$ for some $t = \tilde{t}$. Then $\varepsilon |\lambda(\tilde{t})| = 0(\sqrt{\varepsilon})$ and the interval length diminishes exponentially to $0(\sqrt{\varepsilon})$ when t approaches \tilde{t} . This is completely in accordance with the behavior of the solution of the differential equation.

4. Difference approximations.

In this section we consider the system (3.1)

$$(4.1) \quad dy/dt = A(t)y + F(t), \quad t \geq 0,$$

and approximate it by general difference approximations. We start with multistep methods,

$$(4.2) \quad L[u] \equiv \sum_{j=-1}^r (\alpha_j I + k\beta_j A(t_{v-j}))u_{v-j} = -k \sum_{j=-1}^r \beta_j F(t_{v-j}),$$

where $\alpha_{-1} = 1$ and $\beta_{-1} < 0$. We write (4.2) as a one-step method

$$(4.3) \quad v_{v+1} = \tilde{B}(t_v, k)v_v + k\tilde{F}_v, \quad v = 0, 1, 2, \dots$$

with

$$v_v = \begin{pmatrix} v_v^{(1)} \\ \vdots \\ v_v^{(n(r+1))} \end{pmatrix} = \begin{pmatrix} u_{v+r} \\ \vdots \\ u_v \end{pmatrix}, \quad \tilde{F}_v = (I + k\beta_{-1}A(t_{v+1}))^{-1} \begin{pmatrix} - \sum_{j=-1}^r \beta_j F(t_{v-j}) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\tilde{B}(t, k) = \begin{pmatrix} \tilde{E}_0(t, k) & . & . & . & . & \tilde{E}_r(t, k) \\ I & 0 & . & . & . & 0 \\ 0 & I & 0 & . & . & 0 \\ . & . & . & . & . & . \\ 0 & . & . & 0 & I & 0 \end{pmatrix}$$

where

$$\tilde{E}_j(t, k) = -(I + k\beta_{-1}A(t+k))^{-1}(\alpha_j I + k\beta_j A(t-jk)), \quad j = 0, 1, \dots, r.$$

The aim of this section is to prove theorem 1.3. For this we need the following two well known lemmata.

Lemma 4.1. Consider a class \mathcal{K} of problems and assume that the approximation is stable for this class. Let $A(t)$ be the defining matrices and consider the class $\tilde{\mathcal{K}}$ of problems where $A(t)$ is replaced by $\tilde{A}(t) = A(t) + B(t)$. If the $B(t)$ are uniformly bounded then the approximation is also stable for the class $\tilde{\mathcal{K}}$.

Lemma 4.2. Consider a class \mathcal{K} of problems and assume that there are constants $\eta > 0$, $K > 0$ such that for every problem in \mathcal{K} :

- 1) the interval of integration can be divided into subintervals $t_i \leq t \leq t_{i+1}$ with $t_{i+1} - t_i \geq \eta$
- 2) there is a transformation $T_i(t)$ in each subinterval which satisfies

$$|T_i| |T_i^{-1}| + |dT_i/dt| \leq K$$

- 3) the approximation (4.2) is stable in each subinterval for the transformed problem

$$d\tilde{y}/dt = \tilde{A}\tilde{y}, \quad \tilde{A} = T_i^{-1}AT_i + T_i^{-1}dT_i/dt.$$

If the local stability constants K_I, K_S are uniformly bounded then the approximation is also stable for the class \mathcal{K} .

We make

Assumption 4.1. We consider a class $\mathcal{M}(\Omega) = \mathcal{M}(\Omega, c, \rho, \delta, K)$ (see definition 1.2) for which $c = 0$ and $1 - \delta$ is sufficiently small. Also, the approximation (4.2) satisfies assumption 1.2.

This assumption is no restriction. 1) If $c \neq 0$ then every $A \in \mathcal{M}(\Omega)$ can be written in the form $A = A_1 + B_1$ where the B_1 are uniformly bounded. By lemma 4.1 we can neglect B_1 . 2) Let $A(t_0)$ satisfy the inequalities (1.3a), (1.4a) and (1.4b) with $c = 0$. By lemma 2.2 $|e^{A(t_0)t}| \leq 1$ i.e., the relation (3.2a) is satisfied with $K_0 = 1$, $c = 0$. Furthermore, for every eigenvalue λ of $A(t_0)$, there is an a_{ii} such that

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \leq (1 - \delta) |\operatorname{Re} a_{ii}|.$$

Therefore,

$$\begin{aligned} |\operatorname{Re} \lambda| &\geq \delta |\operatorname{Re} a_{ii}|, \quad |\operatorname{Im} \lambda| \leq |\operatorname{Im} a_{ii}| + (1 - \delta) |\operatorname{Re} a_{ii}| \leq \\ &\leq (\rho + 1 - \delta) |\operatorname{Re} a_{ii}| \leq \delta^{-1} (\rho + 1 - \delta) |\operatorname{Re} \lambda|, \end{aligned}$$

and the relation (3.2b) holds with $c = 0$ if we replace ρ by $\rho_1 = (\rho + 1 - \delta)/\delta$. Thus, if $A(t) \in \mathcal{M}(\Omega, 0, \rho, \delta, K)$ then $A(t_0) \in \mathcal{F}_1(1, 0, \rho_1)$. Furthermore, $A(t) - I$ is negative dominant and, by lemma 2.4, the logarithmic derivative $(A - I)^{-1} dA/dt$ is uniformly bounded. Therefore, by theorem 3.2, we can transform the class $\mathcal{M}(\Omega)$ to another class $\mathcal{M}(\Omega)$ where $1 - \delta$ is as small as we like. The transformation is piecewise constant for every A and $|T_i| |T_i^{-1}|$ is uniformly bounded. By lemma 4.2 we need only consider this new class.

To simplify the notation we shall use

Definition 4.1. $A(t)$ is weakly negative dominant if (1.3a), (1.4a) and (1.4b) hold with $c = 0$.

We shall also simplify the difference approximation. For $c = 0$ the matrices $A(t) - I$, $A(t) \in \mathcal{M}(\Omega)$ are negative dominant and, by lemma 2.5,

$$(4.4) \quad A(t - jk) - I = (A(t) - I)(I + jkC_j(t, k)),$$

where the C_j are uniformly bounded. The matrices

$$(4.5) \quad E_j(t, k) = (I + k\beta_{-1}A(t))^{-1}(\alpha_j I + k\beta_j A(t))$$

are uniformly bounded and have uniformly bounded derivatives. Furthermore, by (4.4),

$$\tilde{E}_j(t, k) = E_j(t, k) + k\tilde{C}_j(t, k)$$

where the $\tilde{C}_j(t, k)$ are uniformly bounded. Therefore, by lemma 4.1, we can replace (4.3) by the difference approximation

$$(4.6) \quad v_{\nu+1} = B(t_\nu, k)v_\nu + k\tilde{F}_\nu, \quad \nu = 0, 1, 2, \dots$$

where B has the same form as \tilde{B} with \tilde{E}_j replaced by E_j .

We consider now some special cases. Let k be fixed and denote by $\mathcal{M}_{0,d}$ the class of problems where A is negative dominant and $|kA| \leq d$ for all t . Here d is a sufficiently small constant. By (4.5)

$$E_j = \alpha_j I + \hat{\beta}_j kA + kAQ_j kA, \quad \hat{\beta}_j = -\alpha_j \beta_{-1} + \beta_j,$$

where Q_j are analytic functions of kA . Therefore we can write the matrix $B(t, k)$ in the form

$$B = B_0 + \hat{A}B_1 + \hat{A}Q\hat{A}$$

where

$$\hat{A} = \hat{A}(t) = \begin{pmatrix} kA(t) & 0 & . & \dots & 0 \\ 0 & kA(t) & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & kA(t) \end{pmatrix},$$

$$B_0 = \begin{pmatrix} \alpha_0 I & \alpha_1 I & \dots & \alpha_r I \\ I & 0 & \dots & 0 \\ . & . & . & . \\ 0 & . & 0 & I & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} \hat{\beta}_0 I & \hat{\beta}_1 I & \dots & \hat{\beta}_r I \\ 0 & . & \dots & 0 \\ . & . & . & . \\ 0 & . & \dots & 0 \end{pmatrix},$$

and $Q(t)$ is an analytic function of kA . By assumption 1.2, the approximation (4.2) is stable in the usual sense. Therefore, there is a nonsingular transformation

$$T = \begin{pmatrix} t_{11} I & \dots & t_{1r+1} I \\ . & . & . \\ t_{r+11} I & \dots & t_{r+1r+1} I \end{pmatrix}$$

such that

$$TB_0 T^{-1} = \begin{pmatrix} D_1 & 0 & . & . & 0 \\ 0 & D_2 & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & D_s \end{pmatrix}$$

where

$$D_j = \kappa_j(0)I, \quad |\kappa_j(0)| = 1, \quad j = 1, 2, \dots, s-1, \quad \kappa_j \neq \kappa_i \quad \text{for } j \neq i,$$

and

$$|D_s| < 1 - \sigma, \quad \sigma = \text{const.} > 0.$$

We can consider the matrix $\hat{A}B_1$ as a perturbation of B_0 . Therefore, we can use the following theorem of B. Engquist [4].

Theorem 4.1. There is a transformation

$$H = T + \hat{A}R$$

where R is an analytic function of kA such that

$$(4.7) \quad HBH^{-1} = D + \hat{A}\hat{Q}\hat{A}.$$

Also, \hat{Q} is an analytic function of kA and

$$(4.8) \quad D = \begin{pmatrix} \kappa_1(0)(I + kg_1A) & 0 & . & \dots & 0 \\ 0 & \kappa_2(0)(I + kg_2A) & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & . & 0 \dots D_s + kg_sA \end{pmatrix}.$$

We can now prove

Lemma 4.3. Consider the class of problems $\mathcal{M}_{0,d}$. For sufficiently small d and k there are constants $\eta > 0$, $K_{1s} > 0$ such that the solutions of (4.6) satisfy the estimate

$$(4.9) \quad |v(t_\nu)| \leq K_{1s} (e^{\eta q(t_\nu)} |v_0| + \max_{0 \leq j \leq \nu} |A^{-1}(t_j)F(t_j)|)$$

with

$$(4.10) \quad q(t_\nu) = \sum_{0 \leq j \leq \nu-1} a(t_j)k, \quad a(t_j) = \max_i \text{Real } a_{ii}(t_j).$$

Proof. Let $w_\nu = H_\nu v_\nu$, then the equation (4.6) can be written as

$$(4.11) \quad H_\nu H_{\nu+1}^{-1} w_{\nu+1} = (D_\nu + \hat{A}_\nu \hat{Q}_\nu \hat{A}_\nu) w_\nu + k H_\nu \tilde{F}_\nu.$$

By lemma 2.5

$$kA(t+k) = kA(t)(I + kC(t,k))$$

where C is uniformly bounded. Therefore

$$H_{\nu} H_{\nu+1}^{-1} = I + k \hat{A}_{\nu} \hat{P}_{\nu}$$

where \hat{P}_{ν} is also uniformly bounded. Furthermore,

$$k H_{\nu} \tilde{F}_{\nu} = \hat{A}_{\nu} G_{\nu}, \quad G_{\nu} = k H_{\nu} \hat{A}_{\nu}^{-1} \tilde{F}_{\nu},$$

with

$$(4.12) \quad |G_{\nu}| \leq \text{const.} \sum_{j=-1}^r |A^{-1}(t_{\nu}) F(t_{\nu-j})| \leq \text{const.} \sum_{j=-1}^r |A^{-1}(t_{\nu-j}) F(t_{\nu-j})|.$$

We first consider the homogenous equation (4.11), i.e., $\tilde{F} = 0$. For sufficiently small d , the row sums $R_1^{1)}$ of the right side of (4.11) can be estimated by

$$\begin{aligned} R_1 &\leq 1 - \frac{1}{2} \sigma, \quad \text{for the rows corresponding to } D_s, \\ R_1 &\leq |1 + g_j k \text{Real } a_{ii} + g_j k \text{Im } a_{ii}| + (1 - \delta) g_j k |\text{Real } a_{ii}| \\ (4.13) \quad &+ \text{const. } d |k \text{Real } a_{ii} + k \text{Im } a_{ii}| \leq \\ &\leq 1 - |g_j k \text{Real } a_{ii}| + \text{const. } d(1 + \delta) |k \text{Real } a_{ii}| \leq \\ &1 - \frac{1}{2} g_j k |\text{Real } a_{ii}| \quad \text{otherwise.} \end{aligned}$$

For the row sums R_2 of the left hand side we obtain, correspondingly

$$R_2 \geq 1 - \text{const. } k^2 |\text{Real } a_{ii}|$$

and the lemma follows from $\tilde{F} = 0$.

Now assume that $w_0 = v_0 = 0$. Consider (4.11) for $0 \leq t_j \leq t_{\nu}$ and let $w_{\mu}^{(\sigma)} = \max_{i,j} |w_j^{(i)}| = \|w\|$. Using row σ of (4.11) with $\nu + 1 = \mu$ we obtain from (4.13)

1) The row sums of a matrix (a_{ij}) are defined by $\sum_{j=1}^n |a_{ij}|$, $i = 1, 2, \dots, n$.

$$(\frac{1}{2} \sigma - 0(k)) \|w\| \leq \text{const. } k |\text{Real } a_{ii}| \|G\| ,$$

or

$$\frac{1}{2} g_j (1 + 0(k)) k |\text{Real } a_{ii}| \|w\| \leq \text{const. } k |\text{Real } a_{ii}| \|G\| ,$$

which proves the lemma.

The next special case is treated in

Lemma 4.4. Let d_1, d_2 be constants with $0 < d_1 < d_2$, and k be fixed. Consider all problems belonging to $\mathcal{M}(\Omega)$ for which the eigenvalues of the matrices kA belong to Ω_{d_1, d_2} (see definition 1.2). Then, there are constants K_{2S} and σ , with $0 < \sigma < 1$, such that the solutions of (4.6) satisfy the estimate

$$(4.14) \quad |v(t_\nu)| \leq K_{2S} \left((\sigma + 0(k))^\nu |v_0| + \max_{0 \leq j \leq \nu} \left| \frac{A^{-1}(t_j) F(t_j)}{1 - \sigma + 0(k)} \right| \right) .$$

Proof. The eigenvalues κ of B have, by assumption, the property that

$$|\kappa| < 1 - \tau .$$

Furthermore, B and dB/dt are uniformly bounded. Therefore, we can without restriction assume that $|B| < 1 - \frac{1}{2} \tau$. Otherwise, we can find a transformation T with $|T| |T^{-1}| + |dT/dt|$ uniformly bounded such that TBT^{-1} has this property. Also,

$$k |\tilde{F}_\nu| \leq \text{const.} \sum_{j=-1}^r |A^{-1}(t_\nu) F(t_{\nu-j})| \leq \text{const.} \sum_{j=-1}^r |A^{-1}(t_{\nu-j}) F(t_{\nu-j})| .$$

Therefore, (4.14) follows from known principles.

We can now prove theorem 1.3 for the case that the approximation (4.2) is strongly stable for the class $\mathcal{N}(\Omega)$ of scalar equations. By lemma 4.2 we need only prove local stability. We consider the homogenous system 4.1 in the neighbourhood of a fixed point t_0 . By lemma 4.1 we can assume that $A(t)$ is negative dominant. Let k be fixed and $d > 0$ a constant. Then we can write $A(t_0)$ in the form

$$A(t_0) = \begin{pmatrix} A_{11}(t_0) & A_{12}(t_0) \\ A_{21}(t_0) & A_{22}(t_0) \end{pmatrix}, \quad |a_{11}| \geq \dots \geq |a_{nn}|,$$

where $A_{22}(t_0)$ is the largest submatrix for which $2|kA_{22}(t_0)| \leq d$. For every eigenvalue λ of A_{11} there is, by the Gershgorin estimates, a diagonal element a_{ii} with

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \leq (1 - \delta) |\operatorname{Re} a_{ii}|, \quad \text{i.e.,} \quad |\lambda| \geq \delta |a_{ii}|.$$

Furthermore, the row sums r_i of A_{11} satisfy the estimate $kr_i > \frac{1}{2}d$. Therefore

$$|ka_{ii}| = kr_i - k \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \geq \frac{1}{2}d - (1 - \delta)|ka_{ii}|.$$

Thus $|ka_{ii}| \geq \frac{1}{4}d$ and the eigenvalues $\lambda(A_{11})$ of A_{11} satisfy the estimate

$$|k\lambda(A_{11})| \geq \frac{1}{4}d\delta.$$

By lemma 2.5, there is a neighbourhood $|t - t_0| \leq \eta$, $\eta > 0$ a constant independent of A and t_0 , such that the matrices $A_{ii}(t)$, $i = 1, 2$

have the property

$$|kA_{22}(t)| \leq d, \quad |k\lambda(A_{11})(t)| \geq d_1, \quad d_1 \geq \frac{1}{8}d\delta.$$

Now consider the approximations (4.6)

$$(4.16) \quad v_{\nu+1}^{[1]} = B^{[1]}(t_\nu, k)v_\nu^{[1]} + k\tilde{F}_\nu^{[1]},$$

$$(4.17) \quad v_{\nu+1}^{[2]} = B^{[2]}(t_\nu, k)v_\nu^{[2]} + k\tilde{F}_\nu^{[2]},$$

for the subsystems

$$(4.18) \quad dy^{[1]}/dt = A_{11}y^{[1]} + F^{[1]},$$

$$(4.19) \quad dy^{[2]}/dt = A_{22}y^{[2]} + F^{[2]},$$

respectively. We choose d so small that (4.17) satisfies the estimate (4.9). If $\delta \rightarrow 1$, then the eigenvalues of A_{11} converge to the eigenvalues λ of A with $|k\lambda| \geq d_1$. By assumption 1.2 these eigenvalues belong to $\Omega_{d_1, \infty}$. Therefore, the eigenvalues κ of $B^{[1]}(t, k)$ satisfy the inequality

$$|\kappa| \leq 1 - \frac{1}{2}\tau \quad \text{for} \quad |t - t_0| \leq \eta$$

provided we have chosen $1 - \delta$ sufficiently small. Thus, the solutions of the system (4.16) satisfy the estimate (4.14).

We can now write the differential equation (4.1) in the form (4.18) and (4.19) with $F^{[1]} = A_{12}y^{[2]}$, $F^{[2]} = A_{21}y^{[1]}$. The difference approximation (4.6) can then be represented by (4.16), (4.17) and, for sufficiently small k , the estimates (4.14) and (4.9) give us

$$\|v^{[1]}\| \leq |v_0^{[1]}| + \text{const.} \|A_{11}^{-1}A_{12}\| \|v^{[2]}\|,$$

$$\|v^{[2]}\| \leq |v_0^{[2]}| + \text{const.} \|A_{22}^{-1}A_{21}\| \|v^{[1]}\|,$$

$$\|u\| = \max_{0 \leq j \leq \nu} |u_j|.$$

Therefore, the approximation is stable provided

$$\|A_{11}^{-1}A_{12}\| \|A_{22}^{-1}A_{21}\| = o((1 - \delta)^2)$$

is sufficiently small. This proves theorem 1.3 when the approximation is strongly stable.

The proof of theorem 1.3 for the case that the approximation is only stable is more complicated. We start again with a special case. Let k be fixed and let $\mathcal{M}_{d, \infty}$ denote the class of problems for which A is negative dominant, A^{-1} is weakly negative dominant (see def. 4.1), and $|(kA)^{-1}| \leq d$. Then we have

Lemma 4.5. Consider the class of problems $\mathcal{M}_{d, \infty}$. For sufficiently small $1 - \delta$, d and k there is a constant $K_{3S} > 0$ such that the solutions of (4.6) satisfy the estimate

$$(4.20) \quad |v(t_\nu)| \leq K_{3S}((1 + kdt_\nu)|v_0| + k(1 + dt_\nu) \max_{0 \leq j \leq \nu} |F(t_j)|).$$

Proof. By (4.5), the $E_j(t, k)$ are now analytic functions of $(kA)^{-1}$ and we can write the matrix $B(t, k)$ in the form

$$B(t, k) = \tilde{B}_0 + \tilde{A}\tilde{B}_1 + \tilde{A}\tilde{Q}\tilde{A}$$

where \tilde{Q} is an analytic function of $(kA)^{-1}$,

$$\tilde{A} = \begin{pmatrix} (kA)^{-1} & 0 & \dots & 0 \\ 0 & (kA)^{-1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \vdots & 0 & \dots & (kA)^{-1} \end{pmatrix},$$

and \tilde{B}_0, \tilde{B}_1 are constant matrices of the same form as B_0, B_1 . By assumption the approximation is stable in a neighbourhood of $\lambda k = \infty$. Therefore, we can apply the same process as used in lemma 4.3. Using theorem 4.1 again we obtain an equation of type (4.11),

$$\tilde{H}_\nu \tilde{H}_{\nu+1} w_{\nu+1} = (\tilde{D}_\nu + \tilde{A}_\nu \tilde{Q}_\nu \tilde{A}_\nu) w_\nu + k \tilde{H}_\nu \tilde{F}_\nu$$

where

$$w_\nu = \tilde{H}_\nu v_\nu, \quad \tilde{H} = T + \tilde{A}\tilde{R}, \quad \tilde{R} \text{ is an analytic function of } (kA)^{-1},$$

and

$$\tilde{D} = \begin{pmatrix} \tilde{\kappa}_1(0)(I + \tilde{g}_1(kA)^{-1}) & \dots & 0 & \dots & 0 \\ 0 & \tilde{\kappa}_2(0)(I + \tilde{g}_2(kA)^{-1}) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & D_s + \tilde{g}_s(kA)^{-1} \end{pmatrix}, \quad \tilde{g}_j > 0, \quad |\tilde{D}_s| < 1 - \sigma.$$

By assumption, A is negative dominant. Thus, by lemma 2.5,

$$(kA(t+k))^{-1} = (I + k\tilde{C})(kA(t))^{-1},$$

and therefore

$$\tilde{H}_\nu \tilde{H}_{\nu+1}^{-1} = I + k\tilde{P}_\nu \tilde{A}_{\nu+1}, \quad \text{with } \tilde{P} \text{ uniformly bounded.}$$

Without restriction we can assume that D only consists of blocks $\tilde{\kappa}_j(0)(I + \tilde{g}_j(kA)^{-1})$ because the block $D_s + \tilde{g}_s(kA)^{-1}$ has no influence

on the form of the estimate. Also, we may assume that all $\kappa_j(0) = 1$.

If not, we can introduce a new variable w by

$$\tilde{w}_\nu = \begin{pmatrix} \kappa_1^\nu & 0 & . & \dots & 0 \\ 0 & \kappa_2^\nu & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & \kappa_{s-1}^\nu \end{pmatrix} w_\nu.$$

Then we can write (4.20) in the form

$$(4.21) \quad (I + k\tilde{P}_\nu \tilde{A}_{\nu+1})w_{\nu+1} = (I + G\tilde{A}_\nu + \tilde{A}_\nu \tilde{Q}_\nu \tilde{A}_\nu)w_\nu + kH_\nu \tilde{F}_\nu,$$

where

$$G = \begin{pmatrix} g_1 I & 0 & . & \dots & 0 \\ 0 & g_2 I & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & g_{s-1} I \end{pmatrix}.$$

We want to estimate $\tilde{w}_\nu = \tilde{A}_\nu w_\nu$. Multiplying (4.21) by \tilde{A}_ν gives us

$$(4.22) \quad (I + k\tilde{A}_\nu \tilde{P}_{\nu+1})\tilde{w}_{\nu+1} = (I + G\tilde{A}_\nu + \tilde{A}_\nu^2 \tilde{Q}_\nu)\tilde{w}_\nu + k\tilde{A}_\nu H_\nu \tilde{F}_\nu, \quad \tilde{w}_0 = \tilde{A}_0 w_0.$$

Equation (4.22) is of the same form as (4.11). (Observe that $\tilde{A}_\nu \tilde{Q}_\nu = \tilde{Q}_\nu \tilde{A}_\nu$.)

Therefore, the estimate (4.9) holds and we obtain

$$|\tilde{w}_\nu| \leq K_{1s}(|\tilde{w}_0| + k \max_{0 \leq j \leq \nu} |H_j \tilde{F}_j|).$$

We can now write equation (4.21) as

$$(4.23) \quad w_{\nu+1} = (I + G\tilde{A}_\nu + \tilde{A}_\nu \tilde{Q}_\nu \tilde{A}_\nu)w_\nu + kH_\nu \tilde{F}_\nu + k^2 \tilde{P}_\nu \tilde{w}_{\nu+1}.$$

The equation (4.23) is again of the form (4.11). We split its solution

into two parts $w_\nu = w_\nu^{[1]} + w_\nu^{[2]}$ where

$$(4.24) \quad w_{\nu+1}^{[1]} = (I + G\tilde{A}_\nu + \tilde{A}_\nu \tilde{Q}_\nu \tilde{A}_\nu) w_\nu^{[1]} + k^2 \tilde{P}_\nu \tilde{w}_{\nu+1}, \quad w_0^{[1]} = w_0,$$

$$(4.25) \quad w_{\nu+1}^{[2]} = (I + G\tilde{A}_\nu + \tilde{A}_\nu \tilde{Q}_\nu \tilde{A}_\nu) w_\nu^{[2]} + k \tilde{A}_\nu H_\nu \tilde{A}_\nu^{-1} \tilde{F}_\nu, \quad w_0^{[2]} = 0.$$

The approximation (4.23) is stable. Therefore we can estimate its solution by

$$|w_\nu^{[1]}| \leq |w_0| + kt \max_{0 \leq j \leq \nu} |\tilde{P}_j \tilde{w}_j| \leq$$

$$|1 + \text{const. } kdt_\nu| |w_0| + \text{const. } kdt_\nu \max_{0 \leq j \leq \nu} |F_j|.$$

(4.9) gives us, for (4.25),

$$|w_\nu^{[2]}| \leq \text{const. } k \max_{0 \leq j \leq \nu} |\tilde{A}_j^{-1} \tilde{F}_j| \leq \text{const. } k \max_{0 \leq j \leq \nu} |F_j|$$

and the estimate (4.20) follows.

We now consider the full system 4.1. Corresponding to our earlier procedure, we can write the matrix A in the form

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}, \quad |a_{11}| \geq |a_{22}| \geq \dots \geq |a_{nn}|,$$

in a neighbourhood $|t - t_0| \leq \eta$, $\eta > 0$ of every point t_0 , where A_{33} , A_{11} are the largest submatrices with $|kA_{33}| \leq d_3$ and $|(kA_{11})^{-1}| \leq d_1$.

Then there are constants d_{21}, d_{22} such that

$$(4.26) \quad d_{21} \leq k |\lambda(A_{22})| \leq k |A_{22}| \leq d_{22}, \quad \lambda(A_{22}) \text{ eigenvalues of } A_{22}.$$

Also

$$(4.27) \quad |A_{33}^{-1} A_{3j}| \leq \text{const.} (1 - \delta), \quad |k A_{3j}| \leq |k A_{33}^{-1} A_{3j}| \leq \text{const.} d_3 (1 - \delta), \quad j = 1, 3.$$

$$(4.28) \quad |A_{22}^{-1} A_{2j}| \leq \text{const.} (1 - \delta), \quad |k A_{2j}| \leq \text{const.} d_{22} (1 - \delta), \quad j = 1, 2.$$

Without restriction we can also assume that

$$|k A_{1j}| \leq \text{const.} (1 - \delta), \quad j = 2, 3.$$

Otherwise, we apply the transformation

$$S = \begin{pmatrix} I & S_{12} & S_{13} \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}, \quad S_{1j} = -A_{11}^{-1} A_{1j}, \quad j = 2, 3,$$

and a simple calculation using (4.27) and (4.28) shows that $S^{-1}AS$ has the desired property.

Now consider the approximation (4.6)

$$(4.28) \quad v_{v+1}^{[i]} = B^{[i]}(t_v, h) v_v^{[i]} + k \tilde{F}^{[i]}, \quad i = 1, 2, 3,$$

for the subsystems

$$(4.29) \quad dy^{[i]}/dt = A_{ii} y^{[i]}, \quad i = 1, 2, 3,$$

and assume that the estimates (4.20), (4.14) and (4.9) are valid. The full system can be written in the form (4.29) with

$$F^{[i]} = \sum_{i \neq j} A_{ij} y^{[j]}.$$

Then stability follows in the same way as earlier from the above estimates provided $1 - \delta$ is sufficiently small.

This proves theorem 1.3 for the case that A_{11}^{-1} is weakly negative dominant.

We shall now show that one can always transform A_{11} so that A_{11}^{-1} has the above property. There is one case for which this is immediately true.

Lemma 4.6. Let A be negative dominant and assume that

$$1/\tau \leq |a_{ii}|/|a_{jj}| \leq \tau, \quad i \neq j, \quad i, j = 1, 2, \dots, n,$$

where $\tau > 1$ is some constant. If

$$\tau \frac{\sigma}{1-\sigma} < 1, \quad \sigma = \frac{1-\delta}{\delta} \sqrt{1+\rho^2}.$$

Then A^{-1} is weakly negative dominant with constants

$$\rho_1 \leq \frac{\rho+\sigma}{1+\sigma}, \quad \delta_1 \leq 1 - \tau \frac{\sigma}{1-\sigma}.$$

Proof. Let

$$\Lambda = \begin{pmatrix} a_{11} & 0 & . & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & a_{nn} \end{pmatrix}.$$

Then

$$A^{-1} = (I + \Lambda^{-1}(A - \Lambda))^{-1} \Lambda^{-1} = (I + C)\Lambda^{-1} = D = (d_{ij}).$$

The norm of $C = (c_{ij})$ satisfies

$$|C| \leq \frac{1-\delta}{\delta}.$$

Therefore, we obtain

$$- \operatorname{Re} d_{ii} \geq - \operatorname{Re} \frac{1}{a_{ii}} - \frac{1-\delta}{\delta} \frac{1}{|a_{ii}|} \geq - (\operatorname{Re} \frac{1}{a_{ii}})(1-\sigma)$$

$$|\operatorname{Im} d_{ii}| \leq |\operatorname{Im} \frac{1}{a_{ii}}| + \frac{1-\delta}{\delta |a_{ii}|} \leq |\operatorname{Re} \frac{1}{a_{ii}}| (\rho + \sigma)$$

for the diagonal elements d_{ii} of D . Furthermore, by (4.30),

$$\sum_{\substack{j=1 \\ j \neq i}}^n |d_{ij}| = \sum_{\substack{j=1 \\ j \neq i}}^n |c_{ij}/a_{jj}| \leq (\tau/|a_{ii}|) \sum_{\substack{j=1 \\ j \neq i}}^n |c_{ij}| \leq \tau \sigma |\operatorname{Re} \frac{1}{a_{ii}}| \leq \frac{\tau \sigma}{1-\sigma} |\operatorname{Re} d_{ii}|.$$

This proves the lemma.

If the diagonal elements of A_{11} are all of the same order of magnitude, i.e., τ is bounded, then we can assume that A_{11}^{-1} is weakly negative dominant, because, as we have seen, we can make $1-\delta$ as small as we like.

We now consider the case that A_{11} contains eigenvalues of different orders of magnitude and we want to transform A_{11} into a block form which separates these eigenvalues. We need

Lemma 4.7. Let

$$A = \begin{pmatrix} a_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & a_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \cdot & \cdot & \cdot & \cdot & \cdot & a_{nn} \end{pmatrix}, \quad |a_{11}| \geq |a_{22}| \geq \cdots \geq |a_{nn}|$$

be a weakly negative dominant matrix and assume that there is a constant

$$\tau > 1 \quad \text{such that} \quad |a_{rr}|/|a_{r+1, r+1}| \geq \tau$$

for some r . If $\tau(1-\delta)/(\tau-1)$ is sufficiently small then there is a

transformation S with

$$\max\{|S|, |S^{-1}|\} \leq 1 + \frac{\tau(1-\delta)}{\tau-1} K_1$$

such that

$$(4.31) \quad \tilde{A} = S^{-1}AS = \begin{pmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{pmatrix},$$

$$\tilde{A}_{11} = \begin{pmatrix} \tilde{a}_{11} & \cdot & \cdot & \cdot & \cdot & \tilde{a}_{1r} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \tilde{a}_{r1} & \cdot & \cdot & \cdot & \cdot & \tilde{a}_{rr} \end{pmatrix}, \quad \tilde{A}_{22} = \begin{pmatrix} \tilde{a}_{r+1 \ r+1} & \cdot & \cdot & \cdot & \cdot & \tilde{a}_{r+1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \tilde{a}_{nr+1} & \cdot & \cdot & \cdot & \cdot & \tilde{a}_{nn} \end{pmatrix}.$$

Here the matrices

$$\tilde{A}_{ii} - K_2 \frac{\tau(1-\delta)}{\tau-1}$$

are also weakly negative dominant and K_1, K_2 are universal constants.

Furthermore, the elements of S are analytic functions of

$$(4.32) \quad \frac{a_{pq}}{a_{ii} - a_{jj}} = \frac{a_{pq}}{a_{pp}} \cdot \frac{a_{pp}}{a_{ii} - a_{jj}}, \quad i = 1, 2, \dots, r; j = r+1, \dots, n; p \geq i, p \neq j.$$

Proof. We write A and \tilde{A} in the form

$$A = \Lambda + A - \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} + (1-\delta) \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

$$\tilde{A} = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} + (1-\delta) \begin{pmatrix} \tilde{B}_{11} & 0 \\ 0 & \tilde{B}_{22} \end{pmatrix}$$

where

$$\Lambda_1 = \begin{pmatrix} a_{11} & 0 & . & . & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & 0 & a_{rr} \end{pmatrix}$$

$$\Lambda_2 = \begin{pmatrix} a_{r+1 \ r+1} & 0 & . & . & 0 \\ 0 & a_{r+2 \ r+2} & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & . & . & . & 0 & a_{nn} \end{pmatrix}.$$

For S we make an ansatz

$$S = I + (1 - \delta) \begin{pmatrix} 0 & T_{12} \\ T_{21} & 0 \end{pmatrix}.$$

Then the relation (4.31) is equivalent to:

$$B_{11} + (1 - \delta) B_{12} T_{21} = \tilde{B}_{11}, \quad B_{22} + (1 - \delta) B_{21} T_{12} = \tilde{B}_{22},$$

$$\Lambda_1^{-1} B_{12} + T_{12} - \Lambda_1^{-1} T_{12} \Lambda_2 = (1 - \delta) (\Lambda_1^{-1} T_{12} \tilde{B}_{22} - \Lambda_1^{-1} B_{11} T_{12}),$$

$$B_{21} \Lambda_1^{-1} - T_{21} + \Lambda_2 T_{21} \Lambda_1^{-1} = (1 - \delta) (T_{21} \tilde{B}_{11} \Lambda_1^{-1} - B_{22} T_{21} \Lambda_1^{-1}).$$

Neglecting terms of order $(1 - \delta)$ in the last two equations gives us a linear system

$$\Lambda_1^{-1} B_{12} + T_{12} - \Lambda_1^{-1} T_{12} \Lambda_2 = 0$$

$$B_{21} \Lambda_1^{-1} - T_{21} + \Lambda_2 T_{21} \Lambda_1^{-1} = 0$$

which has a unique solution. The elements of T_{ij} are of the form (4.32) and its row sums are bounded by $\tau/(\tau - 1)$. Therefore the lemma follows easily by iteration.

We can now complete the proof of theorem 1.3. Consider the matrix

$$A_{11}(t) = \begin{pmatrix} a_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & a_{1r} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{r1} & \cdot & \cdot & \cdot & \cdot & \cdot & a_{rr} \end{pmatrix}$$

for a fixed point $t = t_0$. Assume that the a_{ii} are ordered such that

$$|a_{11}(t_0)| \geq |a_{22}(t_0)| \geq \dots \geq |a_{rr}(t_0)|.$$

Let $\tau_1^{1/r} > 1$ be a constant. We divide the a_{ii} into classes η_j by the following procedure.

- 1) $a_{11} \in \eta_1$.
- 2) $a_{jj} \in \eta_1$ if $a_{j-1, j-1} \in \eta_1$ and $|a_{j-1, j-1}/a_{jj}| < \tau_1^{1/r}$.
- 3) If η_1 does not contain all eigenvalues then there is a last $a_{p-1, p-1} \in \eta_1$ with $a_{pp} \notin \eta_1$. η_2 is then defined by the first two rules with a_{11} replaced by a_{pp} .

This division of the diagonal elements of A_{11} into classes at $t = t_0$ defines a division into classes in a whole neighbourhood $|t - t_0| \leq \eta$ of t_0 , i.e., $a_{jj}(t) \in \eta_i$ if $a_{jj}(t_0) \in \eta_i$. By assumption $|a_{jj}^{-1} da_{jj}/dt|$ are uniformly bounded. If we choose η sufficiently small then the a_{jj} belonging to the same class satisfy the condition of lemma 4.6.

Those belonging to different classes satisfy the condition of lemma 4.7 provided $1 - \delta$ is sufficiently small. Repeated use of lemma 4.7 shows that we can transform A_{11} to blockdiagonal form where every block satisfies the conditions of lemma 4.6. Furthermore, the transformation S is such that $|S||S^{-1}| + |dS/dt|$ is uniformly bounded. Therefore theorem 1.3 follows from lemmata 4.1 and 4.2 and the earlier stability results.

No new difficulties arise if we consider approximations of type (1.29). We can again split the matrix A into three parts and it is not difficult to prove analogs of the lemmata 4.3 - 4.5. Therefore, theorem 1.3 also holds for approximations of type (1.29).

REFERENCES

1. Dahlquist, G. G.: "Stability questions for some numerical methods for ordinary differential equations", pp. 147-158 in AMS Symp. Appl. Math. (1963).
2. Dahlquist, G. G.: "The sets of smooth solutions of differential and difference equations", pp. 67-81 in Stiff Differential Systems, Willoughby (1974).
3. Ehle, B. L.: "High order A-stable methods for the numerical solution of systems of differential equations", BIT 8 (1968), pp. 276-278.
4. Engquist, B.: "On difference equations approximating linear ordinary differential equations", Report No. 21, Uppsala University, Department of Comp. Science (1969).
5. Gear, C. W.: "Numerical Initial Value Problems in Ordinary Differential Equations", Prentice-Hall, Englewood Cliffs, New Jersey, 1971.
6. Kreiss, H. O.: "Über Matrizen die Beschränkte Halbgruppen erzeugen", Math. Scand. 7 (1959), pp. 71-80.
7. Nørsett, S. P.: "One-step methods of Hermite type for numerical integration of stiff systems", BIT 14 (1974), pp. 63-77.
8. Widlund, O. B.: "A note on unconditionally stable linear multistep methods", BIT 7 (1967), pp. 65-70.
9. Willoughby, R. A. (Editor): Stiff differential systems, Plenum Press, New York, 1974.